UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Humberto Fioravante Ferro

OTIMIZAÇÃO DA PREVISÃO DE CARGA ELÉTRICA DE CURTO PRAZO UTILIZANDO CRITÉRIOS DE SIMILARIDADE ENTRE PERFIS DE CONSUMO

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Raul Sidnei Wazlawick, Dr.

Florianópolis, setembro de 2007

OTIMIZAÇÃO DA PREVISÃO DE CARGA ELÉTRICA DE CURTO PRAZO UTILIZANDO CRITÉRIOS DE SIMILARIDADE ENTRE PERFIS DE CONSUMO

Humberto Fioravante Ferro

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistema de Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Banca Examinadora

_____

Rogério Cid Bastos, Dr. Eng.
(Coordenador do Curso)

_____

Raul Sidnei Wazlawick, Dr. Eng. (orientador)

_____

Dayna Maria Bortoluzzi, Dra. Eng.

_____

Pedro Alberto Barbetta, Dr. Eng.

“La scienza è il capitano, e la pratica sono i soldati. Studia prima la scienza e poi seguita la pratica nata da essa scienza. Quelli che s'innamorano della pratica senza scienza, sono come i nocchieri che entrano nella nave senza timone o bussola.”

LEONARDO DA VINCI, Codici

Ao prof. Fioravante Ferro, Dr., por muito ter me ensinado sobre a virtude de se viver.

À profa. Lydia Semenow, Dra., por ter me feito entender que quando algo parece muito difícil, provavelmente está sendo feito do jeito errado.

A D. Maria Elena, ao eng. Paulo Lopes e aos seus filhos Ana, Karina e Cássio, por terem me acolhido entre os seus nos momentos mais difíceis de minha vida.

AGRADECIMENTOS

Ao prof. Wazlawick, Dr., meu orientador, pela paciência e por todos os seus conselhos, além da oportunidade dada junto ao programa de Pós Graduação em Ciência da Computação da Universidade Federal de Santa Catarina. Graças a ele, consegui iniciar e acabar um projeto de pesquisa com sucesso. Simplesmente obrigado, professor!

Ao prof. Rogério Cid Bastos, Dr., pelo apurado senso científico e pelas inúmeras sugestões dadas ao longo desta dissertação. Ele é quase um quiromante da estatística, conseguindo distinguir padrões onde outros só vêem ruído branco. Grato, professor!

Aos integrantes do Projeto PCarga, em especial o Cláudio M. de Oliveira, Dr., e o Luiz Ângelo D. de Luca, pelo tempo dispendido no auxílio à realização deste trabalho. Não fosse por vocês, eu ainda estaria procurando por um tema de pesquisa...

Ao eng. Cássio Guimarães Lopes, PhD., pelo conhecimento enciclopédico cedido sem parcimônia a qualquer momento do dia ou da noite. O Cássio, às vezes, é um pouco dramático, mas se ele não fosse assim, eu jamais teria aprendido que rir da própria sina pode ser um santo remédio. Valeu por tudo, irmãozinho!

Ao prof. Marco Aurélio Benedetti, Dr., por ter me infundido o gosto pelo trabalho científico e por ter sido sempre franco e justo. A felicidade pode ser efêmera, mas também não há mazela que dure para sempre. Um dia chove na horta, Benedetti!

Às Centrais Elétricas de Santa Catarina S.A. - CELESC, pela cessão de seu histórico de medições do sistema SCADA.

Ao Luiz André, amigo de todas as horas, pela sua contribuição a este trabalho. Pois quem há de questionar o valor de todas aquelas horas junto a um balcão, tomando cerveja e discutindo assuntos em nada relacionados a modelos inteligentes híbridos?...

À Susi Helen que, mesmo nas madrugadas em que meus modelos inteligentes não pareciam muito inteligentes e deixavam de convergir, sempre me incentivou e apoiou. Susi, há momentos em que estar próximo representa muito – obrigado por tudo!

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE EQUAÇÕES

ÍNDICE DE TABELAS

RESUMO

As novas condições do mercado de energia elétrica exigem que as concessionárias mantenham altos níveis de qualidade nos seus sistemas. Isso se traduz em aspectos como nível de carregamento, continuidade de fornecimento e, mais recentemente, atuação inteligente no mercado atacadista de energia. Neste contexto, as informações atuais ou históricas não são suficientes para subsidiar o processo decisório: é desejável conhecer as condições do sistema no futuro, em especial a carga elétrica. Dada a complexidade típica dos sistemas elétricos, diversos trabalhos apresentam soluções para a predição de carga que se baseiam em sistemas inteligentes híbridos. Entretanto, embora estes sistemas sejam eficazes e assegurem uma precisão maior do que a obtida com modelos básicos, muitas vezes eles demandam um alto custo computacional. Esse custo torna-se especialmente crítico na determinação da relevância preditiva das variáveis disponíveis. Neste trabalho, é apresentado um método de classificação para os perfis de consumo que otimiza a seleção das variáveis preditoras. Como os perfis de consumo são processos estocásticos variantes no tempo, as técnicas convencionais de extração de características não são eficazes na sua representação, formando padrões inconsistentes. Por esta razão, é apresentada uma nova forma de representação baseada no desempenho de regressores SVM (*Support Vector Machine*) que estimam a carga elétrica em função das diversas variáveis com potencial preditivo. Esta técnica é validada mediante a inspeção do espaço de características gerado, o qual forma agrupamentos que compartilham os mesmos conjuntos de preditoras.

Palavras-chave: previsão de carga, inteligência artificial, extração de características, SVM, redes neuronais.

ABSTRACT

In order to achieve high quality standards in electrical power systems, utility companies rely upon load forecasting to accomplish critical activities such as optimal dynamic dispatch and smart performance in the power wholesale market. Several works propose hybrid intelligent forecasting models to deal with the dynamic and non-linear characteristics of the load at a relatively high computational cost. While such approaches give emphasis to the forecasting itself, this work presents a procedure to detect similarities among distinct consumption profiles. Empirical results show that similar profiles share similar sets of relevant predictors. As finding similarities among profiles is less costly than finding the set of relevant predictors from scratch, a new parameter selection method is proposed. Such method is employed to build some neural forecasters with considerable improvement in the learning time.

Key words: load forecasting, artificial intelligence, feature extraction, SVM, neural networks.

1 INTRODUÇÃO

A desregulamentação da economia vem afetando a sociedade de muitas formas desde o advento da globalização na década de 90. Novas diretrizes econômicas foram definidas e modificaram o mercado de energia, o que conduziu à dissolução de muitos monopólios energéticos históricos ao redor do globo. Como decorrência deste panorama, o preço da energia elétrica aos consumidores finais tem apresentado uma tendência de queda (Iyer et al., 2003).

Embora a forma como os governos conduzem seus mercados de energia possa variar ligeiramente de um país para o outro, observa-se certa propensão de dividir toda a cadeia de fornecimento de energia entre pequenas concessionárias de geração e transmissão. Estas concessionárias utilizam leilões de energia para comprar ou vender energia no mercado atacadista (Iyer et al., 2003). Além disso, elas devem atender certos padrões de qualidade relacionados à estabilidade de fornecimento, níveis de tensão acurados e outros aspectos técnicos (Santoso et al., 2002; Guo et al., 2006).

Para lograr êxito neste cenário competitivo, as concessionárias lançam mão de diversos recursos técnicos, tal como a predição de carga (Iyer et al., 2003; Tao et al., 2004; Guo et al., 2004; Niu et al., 2005). Se a predição não for apropriadamente contemplada, as concessionárias podem sofrer perdas financeiras severas (Guo et al., 2004; Hong et al., 2005; Niu et al., 2005), além de sujeitarem os consumidores a problemas de fornecimento (Santoso et al., 2002). Antever a demanda de energia permite o aprimoramento de muitas atividades críticas para o setor, como a compra de energia, controle de geração, chaveamento de carga, negociação de contratos e planejamento de infraestrutura (Iyer et al., 2003; Tao et al., 2004; Hong et al., 2005; Guo et al., 2006).

O mercado de energia é segmentado em três atividades distintas, mas fortemente acopladas: *geração*, *transmissão* e *distribuição* (Pansini, 2005), mostradas esquematicamente na Figura 1. A geração utiliza usinas de força que, mediante processos eletromagnéticos, produzem eletricidade a partir de outros tipos de energia (térmica ou potencial hidráulica). A transmissão é responsável por estruturas que coletam a energia elétri-

ca das usinas, elevam seu nível de tensão e a transportam até os centros consumidores por meio de *linhas de transmissão*. Nos centros consumidores, a distribuição encarrega-se de conectar as linhas de transmissão a *subestações de distribuição*, as quais energizam os *transformadores de distribuição* através de circuitos de média voltagem denominados *alimentadores*. Finalmente, ainda na distribuição, os consumidores finais são conectados aos transformadores da distribuição. O escopo deste trabalho é orientado à distribuição, principalmente às *regiões de consumo*, realçadas no diagrama da Figura 1.

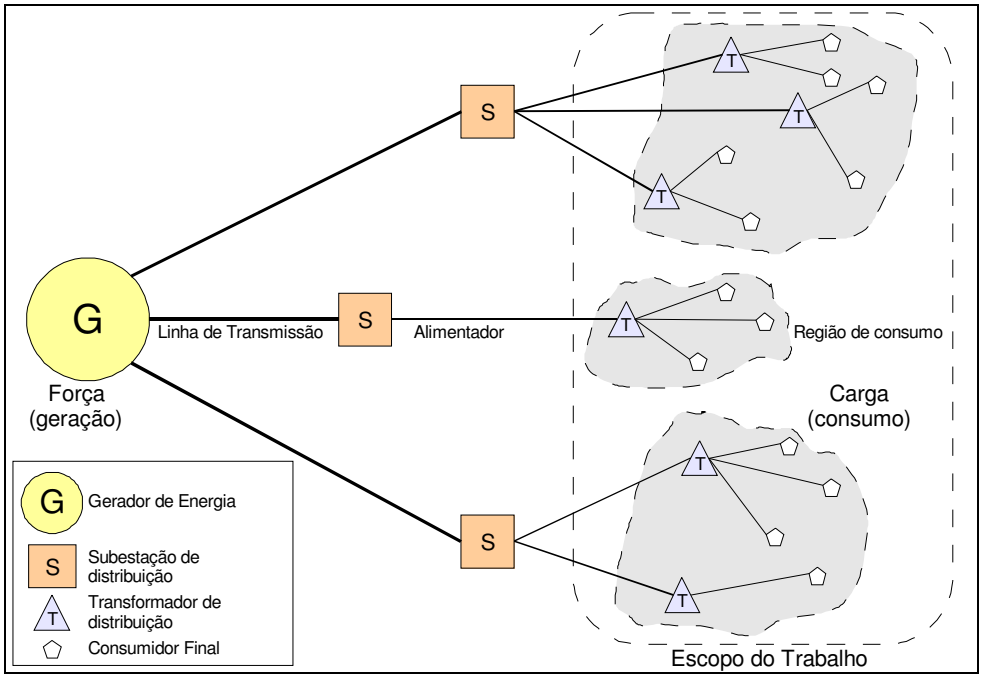

Figura 1 – Diagrama Esquemático do Sistema Elétrico e Escopo desta Dissertação

As subestações abastecem consumidores localizados dentro de regiões que são topologicamente bem definidas (as regiões de consumo) e que apresentam um *perfil de consumo* característico. Estes perfis definem o comportamento da carga elétrica em função das diversas variáveis explanatórias, fornecendo informações valiosas para otimizar a distribuição de energia. Regiões de consumo são estáticas e estabelecem limites geográficos para os perfis; estes últimos, por sua vez, são processos variantes no tempo que descrevem as relações causais entre as variáveis explanatórias e a carga. Um novo perfil de consumo é determinado toda vez que essas relações causais se alteram substancialmente. Assim, as regiões são perenes, contrastando com os perfis, que são transitórios.

A carga mede a potência elétrica que uma região consome num determinado instante. Dependendo do perfil de consumo associado, a carga elétrica pode ser descrita por associações de certas variáveis explanatórias, como temperatura, horário ou situação geográfica. A compreensão destas associações é vital para caracterizar os perfis de consumo, compará-los e construir modelos computacionais capazes de predizer a demanda de energia elétrica.

Este trabalho lida com predição de carga de curto prazo; ou seja, com predição de poucos minutos ou horas à frente (Oliveira, 2004). No entanto, diferentemente de trabalhos correlatos mencionados ao longo desta pesquisa, a proposta desta pesquisa é otimizar algoritmos já existentes. Uma premissa fundamental é que a informação necessária para se construir um modelo de previsão está disponível, restando identificá-la e processá-la de forma adequada. Isso não é uma tarefa trivial, considerando que a previsão de carga elétrica envolve o entendimento de processos não-lineares e variantes no tempo. Portanto, ainda que as variáveis preditoras sejam de fácil obtenção, sua influência na carga pode ser desconhecida ou mal compreendida. Além disso, é possível que o conjunto de variáveis preditoras com efetiva influência na demanda se altere com o perfil sócio-geográfico, conforme demonstrado por Oliveira (2004) e Hong et al. (2005).

Em geral, a construção de um modelo preditor eficiente exige uma escolha criteriosa das variáveis de entrada (Oliveira, 2004; Tao et al., 2004; Hong et al., 2005), as quais tipicamente compõem um subconjunto de todas as variáveis disponíveis, como mostrado na Figura 2. Como destacado nessa figura, o escopo do presente trabalho não consiste nos modelos de predição propriamente ditos, mas na *seleção de parâmetros* e identificação de preditoras (também chamadas de *variáveis explanatórias*, *variáveis preditoras* ou simplesmente *preditoras*[1]).

É relevante observar que o termo *seleção de parâmetros* é empregado na bibliografia de estatística com uma conotação distinta. Neste trabalho, assim como em muitas referências de inteligência artificial e de Teoria do Aprendizado Estatístico, a seleção de

[1] As referências bibliográficas utilizadas para a elaboração deste trabalho são predominantemente escritas em inglês, onde as variáveis preditoras são denominadas de *predictors* (adjetivo), e os modelos de predição, *forecasters* (substantivo). Ambos os termos são traduzidos para o português como *preditores*. Assim, para evitar problemas semânticos, optou-se neste trabalho por traduzir o primeiro como *preditoras* (de *variáveis preditoras*), e o segundo como *preditores* (de *modelos preditores*).

parâmetros consiste no processo de *seleção do modelo* mais apropriado para a predição; ou seja, a seleção de parâmetros define o *conjunto de parâmetros de entrada* que é relevante para a predição [Tao et al., 2004; Hong et al., 2005].

A seleção de parâmetros é necessária para manter a complexidade computacional sob controle e diminuir o tempo de convergência, além de assegurar a precisão das predições. As preditoras constituem os parâmetros fixos do modelo e devem possuir relevância preditiva; caso contrário, o desempenho da predição pode se degradar devido a fatores como ruído ou multicolinearidade (Haykin, 1998; Montgomery et al., 2001). Portanto, a análise preditiva passa a depender também de um mecanismo de seleção de parâmetros.

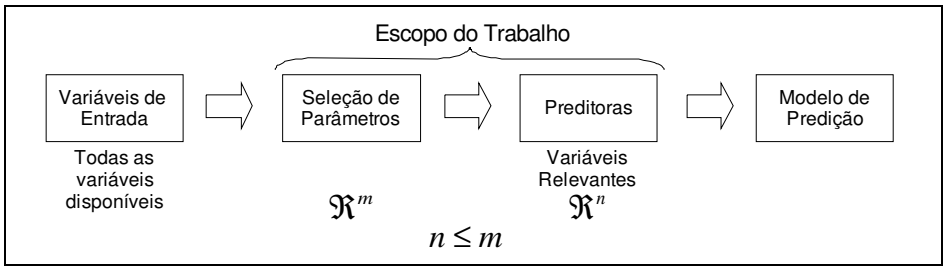

Figura 2 – Esquema da Construção de um Preditor de Carga e Escopo deste Trabalho

É preciso considerar também que o perfil de consumo de uma região é, em geral, dinâmico, o que significa que as relações causais entre os preditores e a carga não é constante. Em outras palavras, a relevância preditiva das variáveis explanatórias pode se modificar ao longo do tempo (Oliveira, 2004); por exemplo, uma dada variável pode ser relevante somente em alguns períodos do ano. Por essa razão, os modelos de predição são transitórios, devendo ser monitorados e, caso necessário, reconstruídos. Sem esta cautela, os erros de predição tendem a aumentar progressivamente – um fenômeno conhecido como *obsolescência* (Oliveira, 2004).

Dentro do cenário descrito, destacam-se três fatores que prejudicam o desempenho da predição de carga:

a) A construção contínua de modelos preditores atualizados, o que se traduz em tempo de processamento.

b) A determinação das variáveis preditoras (parâmetros fixos do modelo[2]).

c) O descarte do conhecimento assimilado pelos modelos obsoletos.

Cada vez que um preditor se torna obsoleto ou um perfil de consumo desconhecido é processado, a seleção de parâmetros deve ser realizada para que um novo preditor possa ser construído. Se iniciada a partir de nenhum conhecimento prévio, tal seleção pode consumir horas de processamento em um *grid* computacional (Oliveira, 2004), desperdiçando o conhecimento incorporado nos modelos já consolidados. Uma forma de minimizar os fatores citados é *criar um critério de similaridade entre os perfis de consumo*, permitindo que os modelos obsoletos subsidiem de alguma forma a construção dos novos modelos.

Este trabalho explora este ponto considerando a seguinte hipótese: *perfis de consumo similares possuem conjuntos similares de preditoras*. Se for demonstrado que esta hipótese é consistente, será possível tornar a seleção de parâmetros mais simples e diminuir o tempo necessário para construir um modelo preditor. Além disso, novos modelos podem ser construídos com base nos já consolidados, dependendo do grau de similaridade entre os perfis de consumo. Por exemplo, se duas regiões de consumo distintas possuem perfis de consumo similares, seus preditores tendem a serem parecidos e empregarão conjuntos de preditoras semelhantes. Quanto maior a similaridade entre os perfis, mais parecidos serão os seus preditores. Teoricamente, no limite, o conjunto de preditoras será o idêntico e mesmo preditor servirá para ambos.

Um outro exemplo é o caso de certas regiões balneárias onde o consumo de energia é alto no verão (alta temporada), mas decresce consideravelmente no restante do ano (baixa temporada). Em tese, na ausência de crescimento vegetativo, a sazonalidade do consumo estabeleceria dois perfis de consumo, um para a alta temporada e outro para a baixa. A Figura 3 mostra este cenário modelado como uma máquina de estados finitos com dois estados, cada um deles correspondendo a uma temporada que, por sua vez, determina um perfil de consumo.

---

[2] Vladmir Vapnik (Vapnik, 1998; Vapnik, 1999) conceitua as variáveis explanatórias como *parâmetros fixos* do modelo em oposição aos *parâmetros ajustáveis*, que são as constantes ajustáveis (fatores de ponderação) utilizadas para combinar linearmente as entradas (variáveis explanatórias) do modelo.

Sempre que uma temporada se encerra, ocorre uma transição. Como existem dois perfis, somente dois modelos são necessários para predizer carga nesta região de consumo. Então, não há necessidade de se criar novos modelos para essa região – tudo que é necessário fazer é alternar o modelo em uso quando ocorre uma transição. Entretanto, a ausência de um critério de similaridade pode fazer com que um novo modelo seja construído toda vez que uma temporada se encerra, simplesmente porque o fenômeno da alternância de perfis não é percebido.

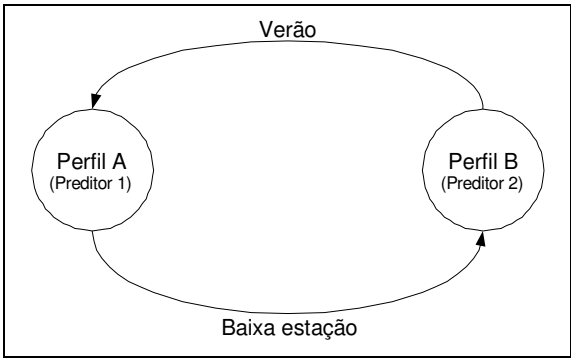

Figura 3 – Alternância de Perfis numa Região de Consumo Hipotética

A Figura 3 mostra um cenário bastante simplificado, onde as vantagens de se reaproveitar o conhecimento existente em preditores consolidados são pequenas. Entretanto, numa aplicação industrial de grande escala onde uma concessionária de distribuição manda instruções de despacho a cada 15 minutos para sua geradora, salvaguardar recursos computacionais pode ser importante.

Não é possível aproveitar o conhecimento acumulado pelos preditores de carga já consolidados, exceto se uma métrica de similaridade entre os perfis for estabelecida. Os esforços deste trabalho estão concentrados na definição de tal métrica, idealizada a partir de dados reais de subestações existentes. Estes dados foram obtidos pelo sistema SCADA das Centrais Elétricas de Santa Catarina (concessionária de distribuição do estado de Santa Catarina).

É preciso notar que a similaridade não se deriva naturalmente dos dados amostrados (brutos), mas requer um processamento elaborado de cada perfil. Como os perfis são na realidade grandes conjuntos amostrados de grandezas elétricas e meteorológicas, um novo método foi criado durante esta pesquisa para extrair características deles. Des-

ta forma, torna-se possível buscar regularidades em um espaço geométrico conveniente (espaço de características), onde os perfis são visualizados como vetores. A mesma abordagem foi modificada para analisar a relevância preditiva das diferentes variáveis em relação à carga, gerando um outro espaço geométrico (espaço causal). Ambos os espaços são descritos neste trabalho e indicam que perfis semelhantes compartilham conjuntos semelhantes de preditoras.

A utilização do método desenvolvido permitiria armazenar os dados de todos os perfis processados até um determinado instante. Após um lapso de tempo razoável, uma base de conhecimento com estas informações poderia minimizar a necessidade de pré-processamento (seleção de parâmetros). De fato, a máquina de estados da Figura 3 mostra que a obsolescência de um preditor ocorre quando o perfil de uma região muda. Se a base de conhecimento for grande o suficiente, toda transição da máquina de estados conduzirá a um perfil conhecido (já processado).

A Figura 4 mostra esse cenário, onde uma região de consumo hipotética evolui através de perfis conhecidos pela base de conhecimento, como indicado pelas setas vermelhas. Os perfis em segundo plano não são atingidos pelas transições dessa região, mas poderiam pertencer a outras regiões. Como os perfis são conhecidos, os preditores associados, bem como os respectivos conjuntos de preditoras, também o são. Desta forma, não há necessidade de se criarem novos preditores e o custo de operação deste sistema se resumiria à recuperação de informações da base de conhecimento.

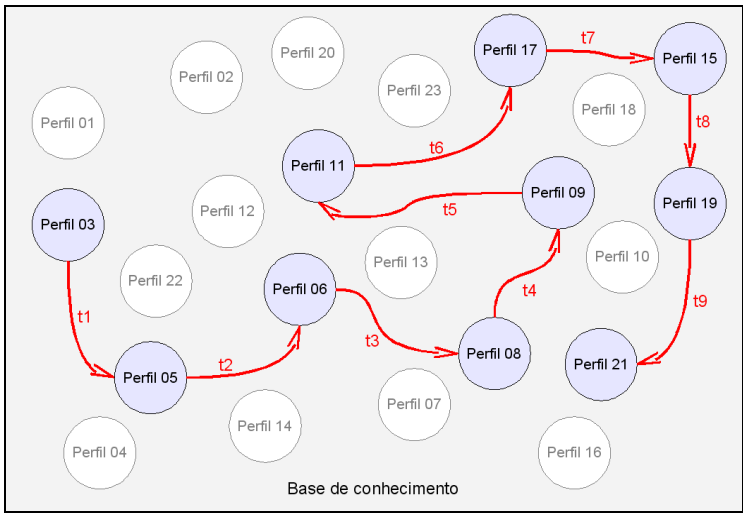

Figura 4 – Evolução dos Perfis de Consumo em uma Região

Ainda que perfis idênticos sejam uma idealidade inalcançável, freqüentemente a similaridade observada é muito grande. Sob estas condições, não se evita a construção de novos preditores, mas o tempo necessário para construí-los é substancialmente reduzido, como será mostrado no Capítulo 6 e discutido no Capítulo 7.

1.1 Objetivos

Conforme salienta a Figura 2, este trabalho não lida diretamente com a predição de carga, mas analisa e propõe mecanismos para otimizá-la computacionalmente. O objetivo geral é identificar padrões úteis na predição de carga dos perfis de consumo. Desta forma, esta pesquisa é orientada a algoritmos de predição como o PCarga (Oliveira, 2004), o SPDS (Guo et al., 2004) ou o GRNN (Iyer et al., 2003), sendo possível que outras abordagens diferenciadas também possam aplicar as técnicas aqui desenvolvidas.

A predição de carga é uma tarefa computacionalmente complexa. De fato, somente a determinação da relevância preditiva das variáveis disponíveis pode consumir horas de processamento, ao passo que os modelos assim gerados podem se defasar rapidamente, não sendo mais capazes de produzir informação útil (Oliveira, 2004). Esta pesquisa explora este aspecto da predição de carga, determinando uma métrica de similaridade para os perfis de consumo tal que perfis semelhantes possuam conjunto de preditoras semelhantes. Com isso, o tempo necessário para descobrir novos preditores de carga é significativamente reduzido, como demonstrado neste trabalho.

1.2 Justificativa

A estrutura do mercado de energia elétrica sofreu transformações radicais em muitos países a partir da década de 90, inclusive no Brasil. A eletricidade passou a ser considerada uma commodity, sendo inclusive utilizada por investidores como opção de investimento cotada em mercados de energia *spot*. Dentro deste novo conceito, os grandes monopólios regionais ou nacionais são pulverizados em unidades menores de gera-

ção e distribuição, as quais integram um grande mercado atacadista de energia (a Câmara de Comercialização de Energia, no caso do Brasil).

Passando ao largo de considerações de cunho ideológico, é possível afirmar que tal conceito é essencialmente benéfico, porque impõe ao sistema elétrico as regras convencionais do mercado de bens e serviço, baseadas nas consagradas relações entre oferta e demanda. Assim, espera-se que o mercado seja capaz de se regular sozinho e que a competição pela venda da commodity no mercado resulte numa queda de preço aos consumidores finais. Do outro lado da relação de consumo, as geradoras e as distribuidoras empenham-se em diminuir os custos operacionais e melhorar a eficiência de suas estruturas físicas, buscando alternativas tecnológicas que subsidiem sua operação e assegurem sua competitividade.

Entre os recursos tecnológicos existentes para assegurar a atuação inteligente do mercado energético, estão as soluções de predição de carga. Antecipar a demanda por energia é vital não somente para os participantes do mercado atacadista de energia, mas também para a sociedade como um todo. A regulamentação prescrita às empresas de energia envolve a manutenção de certos níveis de qualidade que só podem ser mantidos mediante manobras técnicas corretas e investimentos apropriados na infra-estrutura. Por outro lado, os consumidores esperam usufruir o fornecimento independentemente das decisões técnicas ou administrativas necessárias para manter os sistemas de energia em funcionamento. Neste contexto, prever o consumo em horizontes de tempo que variam de poucos minutos a muitos anos a frente é uma vantagem competitiva valiosa.

Embora a predição de energia seja uma área aparentemente consolidada, inclusive com várias soluções comerciais disponíveis, existem alguns aspectos que parecem não estar totalmente sedimentados. Com efeito, grande parte dos trabalhos analisados propõe a aplicação de técnicas inteligentes híbridas baseadas em redes neuronais. Via de regra, estas abordagens consistem em arquiteturas inovadoras, visando sempre aprimorar a precisão obtida, o tempo de convergência ou a diminuição do custo computacional.

No entanto, como a demanda de energia é um processo estocástico dinâmico, os modelos de predição devem ser adaptativos; ou seja, devem acompanhar as mudanças que ocorrem nas regiões de consumo ao longo do tempo (Figura 4). Ora, todos os traba-

lhos analisados são baseados em técnicas conexionistas estáticas, onde a adaptação só é possível durante a fase de aprendizado. Como conseqüência, as predições podem se tornar imprecisas, o que exige a construção de novos preditores.

Esta pesquisa focaliza esse ponto, propondo um método para reciclar o conhecimento existente nos modelos já criados, de forma que eles subsidiem a construção de novos modelos. Com isso, pretende-se diminuir o custo computacional necessário para construir novos preditores, uma vez que isso envolve tempos de processamento relativamente longos. É importante destacar que a redução destes tempos não é meramente uma questão de eficiência computacional, principalmente na previsão de curto prazo (tema desta pesquisa). Em condições reais de operação, o tempo da resposta é tão crítico quanto a sua precisão.

A estratégia delineada neste trabalho consiste em comparar perfis de consumo distintos. O problema inicia na determinação de critérios de similaridade válidos para os perfis, de tal forma que perfis semelhantes utilizem modelos preditores semelhantes. A idéia básica é realizar uma atividade de baixo custo computacional (determinar a similaridade entre os perfis) para poder otimizar uma atividade de alto custo computacional (construir modelos de predição). A otimização é baseada em duas propriedades essenciais da predição de energia, demonstradas ao longo desta pesquisa:

a) Perfis de consumo semelhantes possuem conjuntos semelhantes de variáveis com relevância preditiva.

b) A criação de um preditor para um perfil de consumo desconhecido pode ser facilitada quando iniciada a partir de outro preditor, desde que ambos os preditores estejam associados a perfis de consumo semelhantes.

A propriedade (a) parece expressar uma noção do senso comum, mas a ausência de um critério de similaridade válido impossibilita sua aplicação no contexto da predição de carga. A propriedade (b) é uma decorrência da propriedade (a), uma vez que a relevância dos preditores na carga elétrica constitui fator preponderante na similaridade entre os perfis. Assim, basicamente esta pesquisa consiste em descobrir um critério de

similaridade que satisfaça a propriedade (a) para que possa ser explorada a propriedade (b).

1.3 Metodologia

A literatura de reconhecimento de padrões (por exemplo, Duda 2001 ou Haykin 1994) não aborda a extração de características para sistemas variantes no tempo, dando ênfase às técnicas de reconhecimento propriamente ditas. Outros trabalhos, como o clássico de Oppenheim (1989), apresentam diversas ferramentas matemáticas para processamento de sinais, mas sem apresentar diretamente soluções para a representação de sistemas como os perfis de consumo.

É preciso considerar que, matematicamente, os perfis de consumo são grandes conjuntos amostrados de grandezas meteorológicas e elétricas, normalmente armazenados em estruturas matriciais de alta dimensionalidade. Portanto, não existe uma maneira direta de representá-los convenientemente. Por essa razão, o primeiro passo dessa pesquisa foi criar um espaço de características onde os perfis de consumo pudessem ser visualizados como vetores e uma métrica de similaridade pudesse ser estabelecida. Estabelecido o espaço de características, torna-se necessário classificar os perfis de consumo. Esta etapa foi implementada utilizando escalonamento multidimensional e *k-means*, de forma que perfis semelhantes sejam representados dentro dos mesmos agrupamentos. Após a classificação, o perfil é comparado com outros, obtidos anteriormente. Se a comparação indicar similaridade com algum perfil armazenado, o modelo associado a este último é recuperado e utilizado para construir o novo modelo. Com isso, o processo de seleção de parâmetros pode se tornar mais simples e mais rápido.

A Figura 5 mostra o sistema descrito como um diagrama de blocos. Como mostrado, a extração de características deve ser capaz de representar um determinado perfil de consumo através do vetor de características, constituído por um grupo manuseável de números. Também é mostrado um banco de dados, denominado de *Base de Conhecimento*. Esta base armazena dados relativos a perfis de consumo já processados, ou seja;

vetores de características, conjuntos ótimos de variáveis preditoras e modelos preditores consolidados.

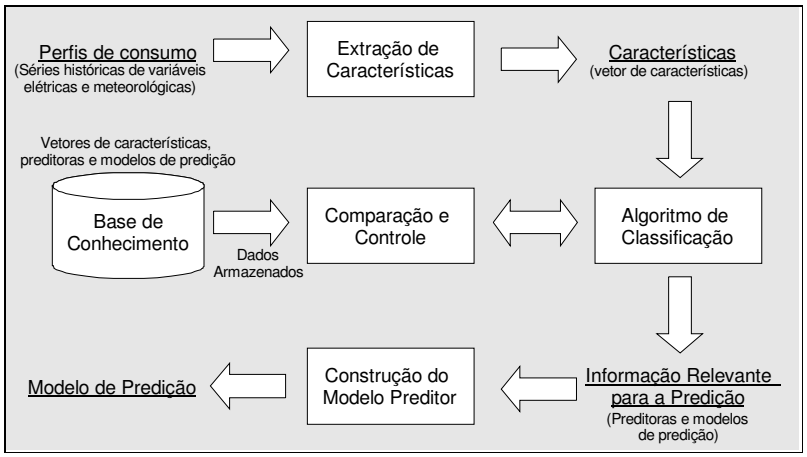

Figura 5 – Sistema de Predição de Carga Otimizado com o Uso de uma Base de Conhecimento

O módulo de *Comparação e Controle* controla o fluxo de informações em toda a estrutura. Este módulo fornece parâmetros para a extração de características e determina qual dos perfis armazenados mais se assemelha com o perfil de entrada. Assim, a construção do novo modelo preditor pode iniciar a partir das informações associadas ao modelo similar. No caso de preditores adaptativos, como as RNAs, isso permite deslocar o ponto de inicialização para uma região próxima do melhor mínimo local conhecido (possivelmente um mínimo global) na superfície de erro, acelerando a convergência.

Dos módulos mencionados, somente a extração de características e a classificação serão descritas neste trabalho. Os demais são omitidos por tratarem de aspectos secundários da pesquisa, tais como o armazenamento de dados e o controle do fluxo de informações.

1.4 Limitações do Trabalho

A solução desenvolvida neste trabalho é essencialmente constituída de uma análise exploratória dos perfis de consumo para determinar regularidades úteis na otimização da predição de carga elétrica. Esta é uma abordagem heurística cuja validade é demonstrada empiricamente para uma série de casos reais, utilizados pelo projeto PCarga (Oli-

veira, 2004). Assim, a regra de similaridade criada é baseada em tendências observadas nos espaços criados, e não necessariamente em demonstrações matemáticas rigorosas.

Em tese, esse fato poderia suscitar dúvidas quanto à generalidade das técnicas desenvolvidas. Não obstante, é oportuno observar que muitas aplicações consolidadas, especialmente aquelas baseadas em modelos conexionistas, fazem uso de abordagens semelhantes. Em todas estas abordagens, sistemas complexos são modelados por estruturas de aprendizado capazes de reproduzir com precisão relações de causa e efeito que, muitas vezes, não são explicadas analiticamente. O mecanismo que assegura esta funcionalidade é a chamada *abordagem empírica* (Vapnik, 1998; Vapnik, 1999), onde as regras causais de um sistema são depreendidas a partir de uma amostra de dados, sem prejuízo da generalidade (*inferência indutiva*).

1.5 Organização do Trabalho

Este trabalho está organizado em sete Capítulos, descritos brevemente a seguir.

O Capítulo 1 apresenta uma breve introdução ao tema, os objetivos, as justificativas, a metodologia desenvolvida e as limitações do trabalho.

No Capítulo 2 são apresentados os conceitos básicos dos sistemas de energia, incluindo algumas breves considerações técnicas, necessárias para a compreensão do cenário em que este trabalho foi desenvolvido.

O Capítulo 3 apresenta uma revisão bibliográfica sobre Estimação e Predição, mencionando elementos fundamentais da Teoria do Aprendizado Estatístico como a minimização de risco.

No Capítulo 4 é apresentada uma revisão bibliográfica sobre Máquinas de Vetores de Suporte, sobre as quais este trabalho se fundamenta.

O Capítulo 5 apresenta sucintamente algumas soluções de predição de carga desenvolvidas recentemente no meio acadêmico.

O Capítulo 6 descreve as hipóteses de trabalho e os procedimentos experimentais levados a cabo para fundamentá-las. Ainda neste Capítulo, são realizadas algumas simulações que demonstram empiricamente a validade do método desenvolvido.

Por fim, o Capítulo 7 apresenta as conclusões e uma discussão final referentes ao trabalho, bem como sugestões de trabalhos futuros.

2 O SISTEMA ELÉTRICO

2.1 Breve Perspectiva Histórica, Definições Essenciais e Alguns Elementos Matemáticos

Para os físicos, todo o Universo é constituído por manifestações de duas entidades fundamentais: massa e energia (Halliday et al., 2004). Massa é a medida da inércia de um corpo, enquanto que energia é a capacidade de um sistema de realizar trabalho.

Todas as atividades humanas dependem de energia. Nos primórdios da humanidade, o ser humano utilizava somente a energia oriunda de seu próprio corpo para atividades como a caça e a manufatura. Com a evolução do conhecimento, surgiram avanços tecnológicos que exigiram novas fontes de energia. Assim, ainda na Antiguidade, surgiram os sistemas mecânicos que dependiam exclusivamente da força motriz de animais ou da energia hidráulica, tais como as rodas d'água e os engenhos.

Os sistemas mecânicos que acompanharam a revolução industrial do século XVIII ficaram maiores e mais complexos. Durante o século XIX, a tecnologia evoluiu e foram concebidos os sistemas a vapor, que propiciaram uma revolução nos transportes com o advento dos motores a vapor, empregados em locomotivas e navios. Muitas das indústrias de então foram se remodelando, ficando mais semelhantes às plantas industriais modernas, aumentando a produção de bens e consumindo mais energia. A partir de então, a questão energética ficou atrelada à economia de um país, tornando-se tão relevante quanto o poderio bélico e, mais do que isso, tornou-se uma métrica importante do grau de desenvolvimento de uma nação. O extrativismo passou a contemplar as novas necessidades energéticas, que demandavam quantidades cada vez maiores de carvão e, mais recentemente, de petróleo. No século XX, o termo matriz energética passou a ser um item importante nos programas de governo das nações civilizadas.

Em meados do século XIX e durante o século XX, as inovações se sucederam com grande velocidade. Surgiram os motores de combustão interna que em muitos casos substituíram as máquinas a vapor (ditas de combustão externa), especialmente nos meios de transporte. Mas os motores de combustão, qualquer que fosse o tipo, apresentavam o mesmo inconveniente dos antigos sistemas hidráulicos: a energia tinha que ser gerada no local onde se precisava dela. Para contornar esta questão de mobilidade, os cientistas contemporâneos, em especial Thomas Alva Edison e Nikola Tesla, determinaram uma forma de transportar energia eficientemente: a *eletricidade* (Meyer, 1971).

A eletricidade é, antes de tudo, uma forma conveniente de transporte de energia. De fato, as aplicações exclusivas para a eletricidade não são tão corriqueiras quanto o senso comum indica; em geral, os dispositivos elétricos ou eletrônicos convertem a eletricidade em outras formas de energia, como luminosa, térmica ou mecânica. Naturalmente, outras formas de transmissão de energia são possíveis – o comitê que administrou a construção da primeira grande hidroelétrica do mundo nas Cataratas do Niágara em 1896, considerou brevemente a utilização de um sistema pneumático antes de se decidir por um sistema elétrico (Meyer, 1971). A idéia era transportar a energia cinética das cataratas até a cidade de Buffalo por meio de ar comprimido. Entretanto, nenhuma outra forma de energia apresenta rendimento tão alto e é tão gerenciável quanto a eletricidade. Provavelmente, se tivesse sido levado a cabo, o grande sistema pneumático do Niágara teria redundado em fracasso.

A geração de eletricidade e os sistemas de distribuição só foram possíveis a partir da descoberta do princípio da indução eletromagnética por Michael Faraday, em 1831 (Halliday et al., 2004). Tal princípio permite o intercâmbio eficiente entre energia elétrica e energia mecânica, sendo aplicado por todas as máquinas elétricas e pelos transformadores de força de qualquer natureza. Todavia, até a idealização dos sistemas de transmissão e distribuição de energia, a descoberta de Faraday não seria de grande aplicabilidade prática.

Os sistemas de transmissão e distribuição de energia foram patenteados por Edison em 1880, um ano depois que ele inventara a lâmpada elétrica (Meyer, 1971). Em 1882, o primeiro sistema comercial de distribuição elétrica foi inaugurado em Manhattam pelo próprio Edison, dando início a um segmento econômico especialmente promis-

sor. A Edison General Electric, fundada por Edison em 1888, ainda existe e tornou-se uma das maiores potências da era Industrial.

Edison construíra um sistema de distribuição de tensão contínua (CC, corrente contínua) trifilar concebido para as lâmpadas elétricas de sua fabricação (Meyer, 1971). Como se tratava de uma implementação comercial, ele dedicou-se a apregoar as vantagens de seu sistema mesmo quando do surgimento de um sistema superior, idealizado por Nikola Tesla em 1887. Isso conduziu a uma grande dissensão entre dois dos maiores nomes da ciência contemporânea, tendo sido registrado na História como a *Guerra das Correntes*.

Tesla, que havia sido colaborador de Edison por alguns anos, propôs que a energia elétrica fosse transmitida e distribuída na forma alternada (CA, corrente alternada) (Meyer, 1971). A tensão alternada pode ser transmitida por grandes distâncias, o que, em princípio, não é viável com a tensão contínua. De fato, os primeiros geradores de Edison não podiam estar a mais de uma milha (aproximadamente 1,6 km) dos consumidores.

O sistema proposto por Tesla é essencialmente o mesmo utilizado hoje em dia no mundo todo. Uma de suas maiores contribuições é permitir que os geradores estejam distantes dos centros consumidores, o que é comum no caso das usinas hidroelétricas.

A primeira grande hidroelétrica do mundo, inaugurada em 1896 nas quedas do Niágara (New York, Estados Unidos), foi projetada pelo próprio Tesla (Meyer, 1971). A usina do Niágara entregava energia elétrica a 26 milhas (42 km) de distância, na cidade de Buffalo (Estados Unidos), um percurso muito maior do que os alcançados pelos sistemas de Edison. Distâncias como essa somente são possíveis porque as perdas de energia no sistema CA são menores do que as observadas em sistemas CC.

As perdas de energia observadas nos sistemas CA são menores porque, depois de gerada, a tensão pode ser elevada a níveis altíssimos, tipicamente centenas de milhares de volts (por exemplo, 138 Kv ou 500 Kv) (Pansini, 2005). Conforme mostra a Equação 1, a potência elétrica P é uma grandeza constante cujo módulo é dado pelo produto da tensão (V) pela corrente (I). Isso significa que, para um mesmo valor de potência, se elevada a tensão, a corrente será rebaixada em razão proporcional.

$$P = V \times I$$

Equação 1 – Potência Elétrica

Com a corrente elétrica sendo mantida em níveis mais baixos devido ao esquema de elevação de tensão, as chamadas *perdas ôhmicas* são reduzidas. Essas perdas causam quedas de tensão na transmissão de energia e superaquecimento dos condutores (efeito Joule) (Halliday et al., 2004).

Uma vez que a resistência do sistema de transmissão é calculada em função da distância, da seção transversal e do material utilizado nos condutores, ela pode ser considerada constante. A resistência deveria ser tão baixa quanto possível; teoricamente, um conjunto de condutores ideal (ou seja, com resistência nula), permitiria a transmissão de uma infinita quantidade de energia. Entretanto, além disso não ser observado em situações práticas, existe uma limitação física na quantidade de energia que pode ser gerada eficientemente por um gerador elétrico, conforme enunciado pelo Teorema da Máxima Transferência de Potência.

A equação 2 mostra a chamada lei de Ohm, que mostra a relação entre a queda de tensão (ΔV) numa linha de transmissão, a corrente e resistência elétricas.

$$\Delta V = R \times I$$

Equação 2 – Lei de Ohm

Ora, se a resistência elétrica R é considerada constante, quanto maior a corrente I que circula pelo sistema de transmissão, maior a queda de tensão observada. Logo, para minimizar esta perda sem reduzir a quantidade de potência elétrica transmitida, é necessário diminuir a intensidade da corrente e, como visto, isto exige o emprego de níveis de tensão maiores. Devido ao princípio da indução eletromagnética, isso só é possível em sistemas AC como os sugeridos por Tesla. Na atualidade, sistemas CC de grande extensão tornaram-se possíveis, mas somente durante a transmissão: a energia é gerada em CA, convertida para CC utilizando dispositivos de estado sólido inexistentes na época de Edison, transmitida em CC, e reconvertida para CA antes da utilização. Esses sistemas, conhecidos como sistemas HVDC (High voltage direct current, corrente contínua de alta tensão), são utilizados em aplicações muito específicas, como cabos de transmissão submarinos.

2.2 Elementos do Sistema Elétrico

O princípio de conservação de energia, enunciado por Lavoisier, atesta a impossibilidade de a energia ser criada. Ela ocorre naturalmente no Universo, podendo ser capturada ou convertida, mas jamais gerada no sentido estrito do termo (Halliday et al., 2004).

Assim, fontes de energia natural devem ser aproveitadas para produzir eletricidade (Pansini, 2005). Em países com muitos recursos hídricos tais como o Brasil, usinas hidroelétricas utilizam a energia potencial da água armazenada em represas ou a energia cinética dos cursos de água para esse fim. Outros tipos de usinas, as termoelétricas, empregam a energia química armazenada em cadeias carbônicas, muito comuns em compostos de origem orgânica como o carvão ou o diesel, liberada através da oxidação. As usinas termonucleares utilizam a energia térmica obtida com a fusão ou, mais recentemente, com a fissão atômica. Também existem formas alternativas de gerar eletricidade, como os geradores eólicos ou solares geralmente utilizados em sistemas de geração distribuída.

Qualquer que seja a forma empregada para tanto, a primeira etapa de um sistema elétrico é a produção de energia elétrica. O termo "geração" é empregado para designar essa etapa ainda que, como ressaltado, se trate de um erro conceitual. Nesse estágio, as tensões são da ordem de 10^3 volts; 20 Kv é um valor típico (Pansini, 2005).

Freqüentemente, os centros consumidores de energia se localizam bem afastados das fontes de produção de energia elétrica; distâncias de centenas ou mesmo milhares de quilômetros são comuns. Para minimizar as perdas ôhmicas, a tensão deve ser elevada por subestações elevadoras a centenas de milhares de volts. Essa etapa é denominada transmissão de energia e emprega valores de tensão como 130 Kv, 230 Kv, 500 Kv ou 750 Kv para percorrer as linhas de força que conectam as usinas geradoras aos centros consumidores (Pansini, 2005).

Nos centros consumidores, as linhas de transmissão são conectadas a subestações rebaixadoras (muitas vezes chamadas simplesmente de *subestações*), responsáveis por diminuir o nível da tensão a patamares tecnicamente mais adequados para o consumo.

Embora as distâncias envolvidas nos centros consumidores sejam menores, a questão da perda ôhmica ainda é um problema. Por essa razão, as subestações entregam energia em níveis de tensão muito mais altos do que os usuários do sistema em geral estão preparados para consumir. Então, as concessionárias de energia empregam um esquema de distribuição que se baseia em dois níveis de tensão, denominadas *tensão primária* (ou MT, média tensão) e *tensão secundária* (ou BT, baixa tensão) (Pansini, 2005).

Os alimentadores são os circuitos que distribuem a tensão primária, utilizando tensões de 5Kv a 23 Kv, sendo 13.8 Kv o valor mais comum no Brasil. Nos pontos geográficos onde os ramais de consumo se concentram (chamados de *centro de carga*), os alimentadores energizam os transformadores de distribuição (TDs) os quais alimentam os ramais dos consumidores finais por meio dos circuitos secundários. Esse esquema destina-se em grande medida aos consumidores residenciais e àqueles cujos consumos estejam abaixo de certos valores padronizados. Grandes consumidores comerciais e industriais são energizados diretamente em tensão primária, responsabilizando-se eles próprios pelo rebaixamento da tensão.

A Figura 6 mostra um diagrama de blocos com todas as etapas de um sistema elétrico moderno. As atividades mais relevantes sob uma perspectiva funcional são mostradas em vermelho, enquanto que as atividades de suporte são mostradas em azul.

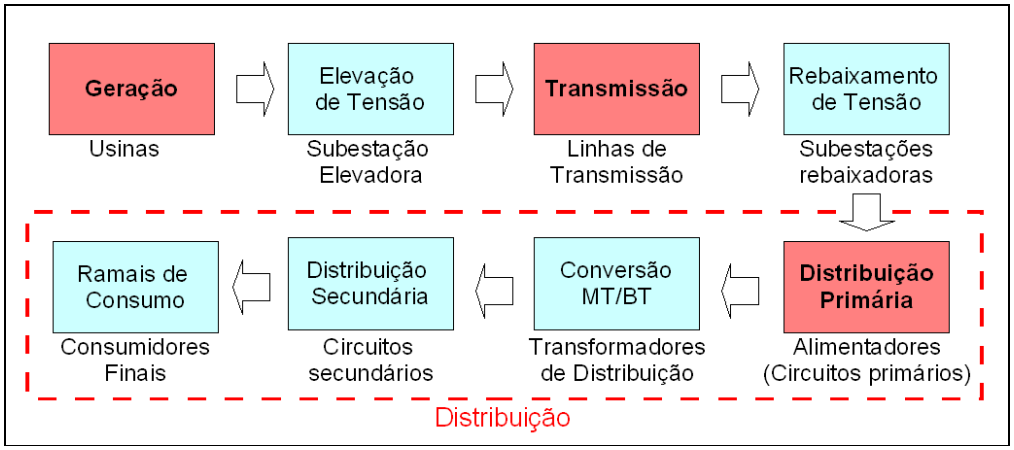

Figura 6 – Elementos de um Sistema Elétrico Moderno

O presente trabalho é voltado à distribuição, cujas etapas estão evidenciadas por uma linha intermitente na Figura 6. No entanto, como os elementos do sistema elétrico

são fortemente acoplados entre si, pode-se aplicar os mesmos conceitos também à geração ou à transmissão. Isso é possível porque o fluxo de potência, da geração ao consumo, é o mesmo; portanto, a predição de energia numa etapa é semelhante à predição em outra etapa.

Em princípio, a diferença entre a transmissão e a distribuição, por exemplo, é um fator de escala: enquanto as linhas de transmissão abastecem alguns poucos centros consumidores, as subestações aplicam uma fração desta energia a centenas de TDs. Evidentemente, é razoável assumir que o perfil de consumo de um TD, por ser mais pulverizado, possa ser mais bem compreendido do que o perfil de uma grande geradora como Itaipu (a maior usina hidroelétrica do mundo), que abastece milhões de consumidores em regiões muito distintas entre si. Entretanto, tanto num caso como no outro, os modelos de predição são não-lineares e dinâmicos, o que requer a adoção de estruturas preditivas igualmente complexas, pelo menos no que tange à previsão de curto prazo.

3 ESTIMAÇÃO E PREDIÇÃO

3.1 Introdução

O conceito de sistema é empregado em muitas áreas da ciência e tecnologia, possuindo uma conotação comum em todas elas: um sistema é *uma entidade cujo comportamento varia em função dos estímulos recebidos* (Oppenheim et al., 1999). Praticamente tudo que nos cerca pode ser considerado como um sistema, por exemplo:

a) Sistemas de distribuição de energia, cujo comportamento (carga elétrica) varia em função dos estímulos (condições atmosféricas, horário, dia da semana, etc).

b) Um automóvel, cujo comportamento (aceleração e velocidade) depende dos estímulos (pressão do acelerador).

c) As commodities do mercado financeiro, cujo comportamento (cotação) depende dos estímulos (políticas fiscais do país, fatores de risco, taxa de inflação, etc).

Muitos sistemas, como os citados, podem ser modelados através das relações causais, também chamadas de *dependências funcionais*. Nestes modelos, os valores de algumas variáveis (estímulos) determinam o valor de outras (comportamento). O estudo e a compreensão dessas dependências podem auxiliar diversas atividades humanas, como o planejamento e o processo decisório nas organizações (Makridakis et al., 1997).

Causalidade é um conceito associado à ânsia humana por padrões. As pessoas dependem de padrões a um grau tal que os consideram pervasivos, o que explica certas compulsões como as teorias da conspiração ou os jogos de azar (Berry et al., 2004). Excetuando-se por essas aplicações espúrias, denominadas de *data dredging* (Hand et al., 2001), a causalidade é evidenciada pela correlação entre eventos (Montgomery et

al., 2001). Em algumas situações, entretanto, os dados precisam ser transformados para que essa correlação seja visualizada apropriadamente, através de técnicas como a linearização ou o logaritmo (Montgomery et al., 2001; Pindyck et al., 2004).

É importante destacar que é possível haver fortes correlações entre alguns eventos sem uma relação causal subjacente (embora o inverso seja sempre verdadeiro) (Berry et al., 2004). Considere-se, por exemplo, o seguinte raciocínio: na cidade de Florianópolis, no verão, a venda de refrigerantes aumenta. Na mesma época, também são observados mais congestionamentos no trânsito e isso estabelece uma correlação entre os dois eventos. Entretanto, é arriscado creditar os engarrafamentos ao consumo de refrigerantes ou o contrário; portanto, não há relação causal plausível entre ambos. É mais provável que a chegada do verão cause um aumento tanto na venda de refrigerantes como no fluxo de motoristas que procuram as praias de Florianópolis, sem que os eventos estejam diretamente relacionados entre si.

3.2 Estimação

Estimação possui conotações distintas na estatística paramétrica e na mineração de dados. Em estatística, o termo estimação é frequentemente utilizado para descrever os procedimentos matemáticos que aproximam (ou seja, *estimam*) os valores dos parâmetros fixos de um modelo estatístico (por exemplo, os coeficientes angulares de um modelo de regressão linear) (Montgomery et al., 2001; Pindyck et al., 2004). Em contrapartida, em mineração de dados o termo está associado aos modelos de regressão propriamente ditos (Berry et al., 2004), enquanto que o procedimento para aproximar os parâmetros de tais modelos é chamado de *aprendizado* (Haykin, 1998; Vapnik, 1998; Vapnik, 1999; Duda et al., 2000). Neste trabalho, foi adotada a conotação empregada na literatura de mineração de dados. Assim, nesta dissertação, *estimação* é a resposta obtida por um modelo de regressão a partir dos valores das variáveis de entrada, enquanto que *estimador* é o próprio modelo de regressão (independentemente de como ele é construído).

Os estimadores descrevem as dependências funcionais de um sistema través de modelos matemáticos convenientes. Estes modelos mostram como os estímulos *X* (também chamados de *entradas*, *variáveis regressoras/explanatórias*, *fatores de influência* ou *preditoras*) influenciam o comportamento *Y* do sistema (respostas) (Montgomery et al., 2001; Hastie et al., 2003; Niu et al., 2005; Guo et al., 2006). Em outras palavras, é possível simular o sistema e prever a sua resposta quando as regressoras assumem determinados valores. Este tipo de simulação é muito conveniente para inúmeros ramos do conhecimento, tais como engenharia, medicina e economia.

O problema de estimação é considerado um problema de inferência estatística, cujo objetivo é formalmente descrito como inferir as dependências funcionais de um sistema através de uma amostra empírica de dados. Os primeiros trabalhos de inferência tinham como base os modelos probabilísticos (funções de distribuições de probabilidade) empregados na descrição de muitos sistemas reais (Vapnik, 1998).

Os principais modelos de inferência estatística foram unificados por Fisher dentro da estatística paramétrica. Com isso, a questão de estimar funções a partir de um conjunto de dados (análise de discriminantes, análise de regressão e estimação de densidade) seria descrita como um problema de estimar os parâmetros de modelos probabilísticos (*paramétricos*) específicos (Vapnik, 1998). Fisher sugeriu ainda um método para estimar os parâmetros desconhecidos em todos esses modelos – o método da máxima verossimilhança (*maximum likelihood method*) (Vapnik, 1998).

As abordagens da Teoria de Aprendizado Estatístico (TAE), em contrapartida, empregam funções determinísticas tais como funções preditivas ou análise de agrupamentos (Herbrich, 2001). Por esta razão, muitos pesquisadores têm observado que as idéias gerais da TAE se assemelham à estimação não-paramétrica que, ao contrário da estatística paramétrica, não considera um modelo probabilístico (distribuição de probabilidade) para o sistema sob estudo (Herbrich, 2001).

O aprendizado estatístico se limita a descrever dados novos que não foram utilizados durante a construção do modelo. A estatística paramétrica vai além disso, ao determinar qual distribuição de probabilidade descreve o sistema. Essa pode ser uma meta ambiciosa, porque tanto a quantidade de dados como a de conhecimento *a priori* nem

sempre é suficiente para tanto (Vapnik, 1998; Herbrich, 2001). Ademais, é possível demonstrar que nem sempre problemas reais podem ser descritos por modelos probabilísticos clássicos (Vapnik, 1998). Isso favorece a aplicação da TAE em muitos problemas reais, porque não requer grandes amostras de dados e dispensa qualquer conhecimento a priori do sistema em questão (Haykin, 1998; Vapnik, 1998).

Ao desprezar o modelo estatístico subjacente a um sistema, a TAE incorre numa limitação semântica: a descrição dos dados gerada pelo modelo não possui interpretação tão clara como a fornecida pelos modelos estatísticos. Um modelo de regressão linear, por exemplo, é definido em função de dois parâmetros, mostrados na Equação 3.

$$y = y(x) = a \cdot x + b$$

Equação 3 – Um Modelo de Regressão Simples

O parâmetro *a* (coeficiente angular) define como a resposta *y* do sistema varia em função do regressor *x*, enquanto que o parâmetro *b* (coeficiente linear) mostra a resposta *y* do sistema quando *x* é nulo (Montgomery et al., 2001).

Por outro lado, o aprendizado estatístico considera conjuntos aninhados de funções implementadas por uma máquina de aprendizado, responsável também por definir uma medida da *capacidade* (grau de complexidade) para cada um destes conjuntos (Vapnik, 1998). Não existe nenhum modelo analítico em particular associado ao estimador gerado: as funções vão sendo testadas iterativamente até que uma delas consiga descrever os dados com precisão suficiente. Conseqüentemente, os parâmetros do estimador gerado serão necessariamente obscuros; ou seja, não terão uma interpretação clara.

Neste trabalho, as possíveis interpretações semânticas que possam ser extraídas dos modelos de estimação não auxiliam na solução do problema proposto. Por si só, tal fato não impede a aplicação dos métodos paramétricos; porém, não existe informação *a priori* suficiente sobre os modelos probabilísticos que descrevem um sistema de distribuição de energia. Ao contrário, os trabalhos de pesquisa nesta área comprovam o caráter dinâmico da carga elétrica (Tao et al., 2004; Guo et al., 2004; Oliveira, 2004; Hong et al., 2005). Isso viola a chamada *premissa da continuidade* (Makridakis et al., 1997) e

torna os parâmetros de um possível modelo probabilístico funções do tempo, inviabilizando a abordagem. Assim, a opção natural para implementar as soluções sugeridas neste trabalho passa a ser a abordagem não-paramétrica.

A definição de sistema dinâmico é intuitiva – um sistema é dinâmico quando suas propriedades variam com o tempo, razão pela qual esse tipo de sistema também é denominado de *sistema variante no tempo*. Matematicamente, isso corresponde à família de equações que são funções explícitas do tempo (Oppenheim et al., 1999). A Figura 7 ilustra graficamente este conceito.

O sinal da Figura 7a é um sinal sinusoidal na forma $y = sen(t)$, enquanto que o sinal da Figura 7b é uma sinusóide degenerada na forma $y = t \cdot sen(t)$ (notar que, no último caso, o termo *t* aparece fora do argumento da função, caracterizando uma dependência explícita do tempo). Se, por hipótese, ambos os sinais forem modelados como processos gaussianos (abordagem paramétrica), eles teriam que ser completamente caracterizados por dois parâmetros: média e variância (Montgomery et al., 2001; Johnson et al., 2002; Pindyck et al., 2004). Isso parece ser adequado para o sinal da Figura 7a, que apresenta ambos os parâmetros constantes, mas não é o caso do sinal da Figura 7b, como discutido a seguir.

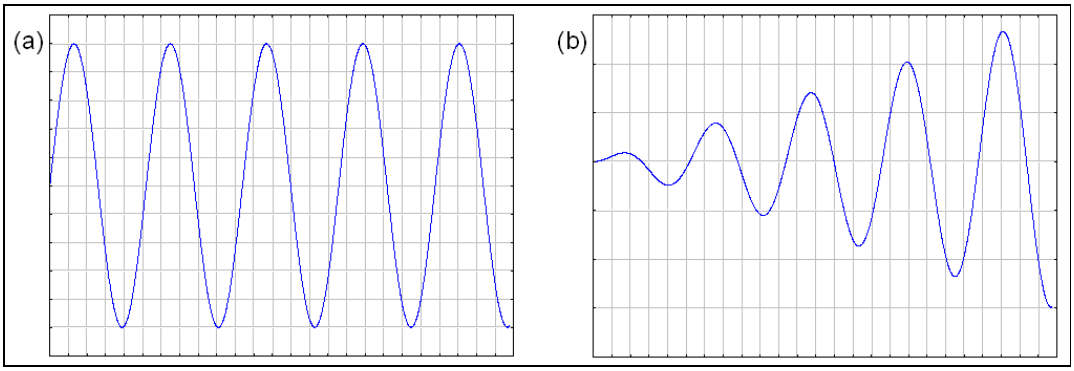

Figura 7 – Variância no Tempo: (a) Sistema Invariante e (b) Sistema Variante

Sejam os gráficos da Figura 8a e da Figura 8b, que mostram a média no tempo para, respectivamente, os sinais mostrados na Figura 7a e Figura 7b. Essas médias são calculadas de forma incremental; ou seja, consideram conjunto de valores progressivamente maiores. Isso significa que, para uma janela arbitrária de *n* amostras, cada curva

da Figura 8 é calculada da seguinte maneira: o primeiro ponto é a média de n amostras, o segundo de $2\times n$ amostras, e assim por diante. Com isso, é possível visualizar como a média se comporta ao longo do tempo e realizar algumas inferências úteis.

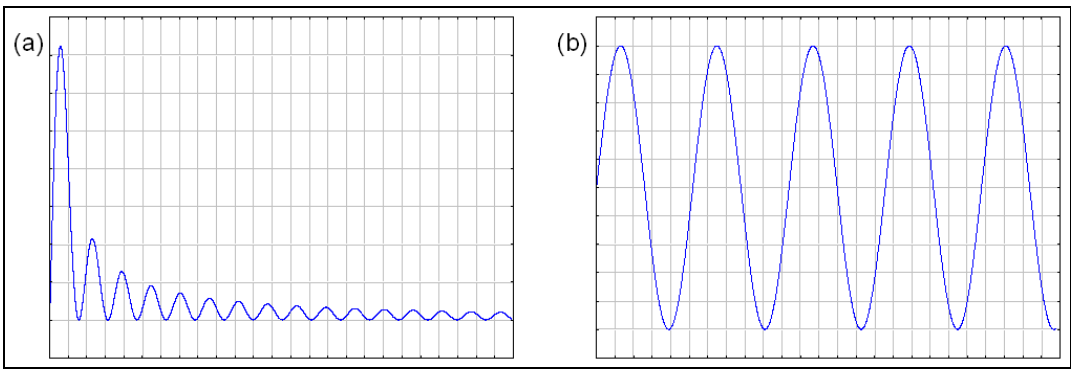

Figura 8 – Médias de Dois Sistemas: (a) Sistema Estacionário e (b) Sistema Não-Estacionário

Como se pode notar, a média do sistema invariante (Figura 8a) tende a um valor definido, enquanto que a do sistema variante (Figura 8b) fica oscilando, sem uma tendência de convergência. A variância apresenta um comportamento geometricamente idêntico para os dois sinais. Naturalmente, o exemplo citado é simplista e, possivelmente, poderia ser contornado para permitir o uso da estimação paramétrica. Entretanto, a carga elétrica e as variáveis preditoras (tais como grandezas atmosféricas) são intrinsecamente mais complexas, conforme discutido no Capítulo 6. Claramente, neste trabalho, é mais interessante criar uma função que reproduza a relação causal entre preditoras e a carga elétrica sem tentar determinar o modelo estatístico subjacente do que criar este modelo e determinar os parâmetros que os caracterizam.

Existem ainda outros fatores que influenciaram a escolha das abordagens não-paramétricas neste trabalho em detrimento das paramétricas. Vladimir Vapnik (Vapnik, 1998) menciona que a abordagem não-paramétrica, desde o seu advento com as redes neuronais na década de 60, capturava *algo* nos problemas reais de alta dimensionalidade que as técnicas descritivas clássicas não conseguiam, senão em situações de baixa dimensionalidade. Este fenômeno tornou-se conhecido a *maldição da dimensionalidade* (Vapnik, 1998; Vapnik, 1999; Herbrich, 2001; Hastie et al., 2003). Por essa razão, a literatura enquadra os métodos não-paramétricos como uma generalização dos métodos paramétricos (Vapnik, 1998). Isso é intuitivo se for considerado que, apesar dos mode-

los paramétricos apresentarem um desempenho restrito em domínios de alta dimensionalidade, os não-paramétricos tratam tanto os domínios de alta como os de baixa dimensionalidade com acentuada desenvoltura.

Um estimador é concebido como o modelo matemático de um sistema; ou seja, ele é uma abstração matemática, capaz de descrever a resposta deste sistema a um determinado conjunto de estímulos (valores de entrada). Esse modelo assume a forma de uma função ou, mais genericamente, de um operador não-linear, ressaltando-se que a linearidade é uma forma específica de não-linearidade (Steinbruch, 1987 et al.). A Equação 4 é uma generalização da Equação 3, que mostra um operador não-linear *f* mapeando o domínio *X* (regressoras) na imagem *Y* (resposta) de um dado sistema. Em situações práticas, admitem-se muitas formas diferentes para *f*, sendo que todas elas admitem um erro residual; ou seja, a modelagem nunca é perfeita (Haykin, 1998; Montgomery et al., 2001).

$$f : X \rightarrow Y$$

Equação 4 – O Operador *f* como um Modelo de Estimação

Em sentido matemático mais amplo, tanto *X* como *Y* são espaços vetoriais; portanto, todo $x \in X$ e $y \in Y$ são vetores (Steinbruch, 1987 et al.). Tipicamente, define-se $X \subset \Re^p$ (onde $p > 0 \in \mathrm{N}$) e $Y \in \Re$ (Herbrich, 2001; Hastie et al., 2003).

A Teoria de Aprendizado Estatístico denomina o processo de criação dos modelos de regressão como *aprendizado supervisionado* (Vapnik, 1999; Hastie et al., 2003). O termo *supervisionado* é empregado porque o operador *f* é determinado a partir de um modelo de aprendizado iterativo. Em cada iteração, a saída $\hat{y}$ estimada por *f* é comparada com *y*, a saída original do sistema. Esse modelo de aprendizado é composto de três componentes, conforme mostra a Figura 9 (Vapnik, 1998; Vapnik, 1999):

a) Um gerador G, que produz vetores aleatórios $x \in X \subset \Re^p$, os quais se distribuem de acordo com uma função de distribuição de probabilidade *F(x)* desconhecida.

b) Um sistema S, que retorna uma saída $y \in Y \subset \Re$ para cada *x* apresentado em sua entrada de acordo com uma função de distribuição condicional *F(y|x)* desconhecida.

c) Uma máquina de aprendizado LM (*Learning Machine*), capaz de implementar um conjunto de funções na forma $\hat{y} = f(x, \alpha)$, onde $\hat{y}$ é uma estimativa de y produzida por f e α é um conjunto de parâmetros ajustáveis, único para cada função.

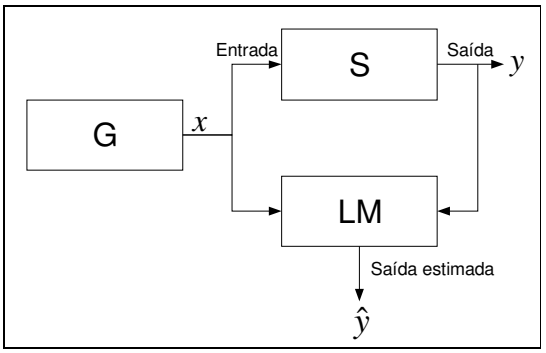

Figura 9 – Modelo de Aprendizado Supervisionado

O conjunto α é que retém o conhecimento de cada modelo gerado; ou seja, é a partir dele que as dependências funcionais são explicadas (Pindyck et al., 2004). No caso do modelo linear simples, mostrado na Equação 3, α é constituído pelos parâmetros *a* e *b*; em se tratando de um modelo conexionista, é o conjunto de pesos sinápticos (Haykin, 1998; Duda et al., 2000). Depois do processo de aprendizado, o conjunto de parâmetros do estimador é mantido constante, razão pela qual é comum substituir-se a notação $f(x, \alpha)$ (dita *notação paramétrica*[3]) por *f(x)* simplesmente.

O aprendizado pode ser descrito como o ato de determinar qual função $f(x, \alpha)$ produz a melhor estimativa para a resposta y do sistema (Vapnik, 1999). Em princípio, quanto menor a diferença entre a saída estimada $\hat{y}$ e a resposta *y*, melhor é o modelo.

[3] O termo *notação paramétrica* é referência aos parâmetros ajustáveis de uma família de estimadores, não se relacionando a *estimação paramétrica* da estatística descritiva. De fato, tanto os métodos paramétricos como os não-paramétricos empregam estimadores com parâmetros ajustáveis.

A Figura 9 mostra ainda um gerador G que produz as entradas x. Esse gerador é uma abstração do processo de amostragem (ou coleta), que seleciona os dados que compõem o conjunto de treinamento. A Equação 5 mostra este conjunto como sendo composto por *L* pares ordenados (*x*, *y*), de forma que cada par associa uma entrada x à saída correspondente *y*. Após o aprendizado, espera-se que o modelo produza respostas acuradas para qualquer elemento de *X*, incluindo as que foram omitidas no conjunto de treinamento. Estatisticamente, isso é possível porque a amostragem representa a população subjacente que representa a verdadeira relação sob estudo (Pindyck et al., 2004).

$$S = \{(x_l, y_l) \in X \times Y\}_{l=1}^{L}$$

Equação 5 – Conjunto de Treinamento

Uma parte dos problemas de modelagem acontece quando o conjunto de treinamento é inadequado, fazendo com que o modelo seja tendencioso (Montgomery et al., 2001; Pindyck et al., 2004). Para evitar isso, o conjunto deve ser *balanceado* (Berry et al., 2004). O balanceamento pode ser entendido informalmente como um método homogêneo de amostragem, de tal sorte que o conjunto de treinamento, embora limitado, seja representativo. Isso pode ser assegurado se os dados do conjunto forem obtidos de forma idêntica e independentemente distribuída (as chamadas condições *i.i.d., independent and identically distributed*) (Vapnik, 1999).

Se o conjunto não puder ser amostrado corretamente, ainda é possível balancear o conjunto de treinamento mediante o uso de fatores de ponderação (Berry et al., 2004). Este procedimento permite ajustar a contribuição de um grupo de amostras no resultado da análise, ponderando seus valores.

3.3 Predição

A *predição* (ou *previsão*) é essencialmente uma espécie de estimação, tal como definida no início da seção 3.2 (Haykin, 1998; Berry et al., 2004). A diferença é que estimador f (Equação 4) mapeia as resposta *y* num instante posterior à amostragem de x. Isso significa que, dado uma entrada *x* obtida num instante t, f retorna uma estimativa para y (ou seja, $\hat{y}$) em $t+\delta$, onde δ é o intervalo de tempo desejado.

Uma das diferenças construtivas mais significativas entre os modelos de estimação e de predição está na forma de compor os conjuntos de dados para o aprendizado do modelo. A Figura 10 mostra de forma esquemática a estratégia normalmente adotada para criar o conjunto de treinamento para um sistema de predição. A idéia é que o modelo possa aprender como os dados do passado remoto afetaram a resposta do sistema no passado recente. Assim, baseando-se na premissa de que o futuro imita o passado, torna-se possível prever o futuro alimentando-se o estimador com dados do passado (Makridakis et al., 1997; Berry et al., 2004).

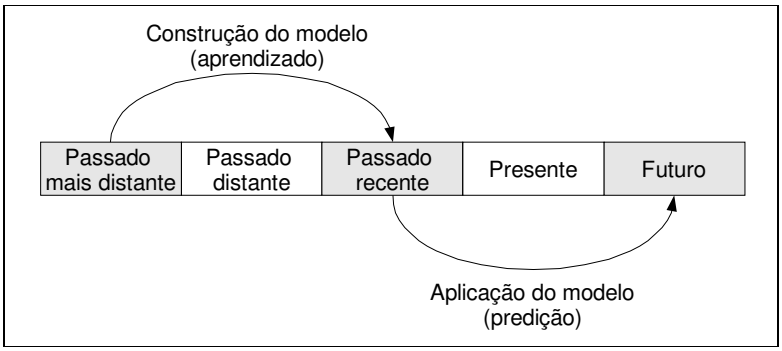

Figura 10 – Dados para um Modelo de Predição

Existem dois tipos básicos de modelos preditores: *explanatórios* e *temporais* (Makridakis et al., 1997).

Os modelos explanatórios são estruturalmente idênticos aos estimadores – a diferença reside no intervalo de tempo entre a amostragem das regressoras e a observação da saída. Durante o aprendizado, f deve prever y no instante $t = k'$ a partir de x amostrado em $t = -k$, onde $-k$ é um instante de tempo anterior à k' (ou seja, $-k < k'$). A Equação 6 mostra uma representação matemática para o modelo descrito, enquanto que a Equação 7 mostra como é representado o conjunto utilizado para o aprendizado desse modelo.

$$y^{k'} = f(x^{-k})$$

Equação 6 – Modelo Preditor Explanatório

$$S = (x_i^{-k_i}, y_i^{k_i'}) \in X \times Y \quad i = 1,2,...,l$$

Equação 7 – Conjunto de Treinamento para um Modelo Explanatório

Assim como no caso dos estimadores, os modelos explanatórios assumem que existe uma dependência funcional entre *X* e *Y*, mas nos moldes da Equação 6. Para que o modelo seja estável, assume-se que a dependência funcional não se altera com o tempo, o que é chamado de *presunção da continuidade* (Makridakis et al., 1997).

Os modelos de séries temporais diferem dos explanatórios por não considerar a dependência funcional mostrada na Equação 4. Ao invés disso, o sistema é considerado uma caixa preta, e a predição é baseada somente nos valores passados (séries históricas) da resposta[4] (Makridakis et al., 1997; Berry et al., 2004). A denominação *temporal* se deve justamente ao emprego de séries históricas, também denominadas de séries temporais.

A Equação 8 modela analiticamente um preditor temporal *f*, representando-o como um mapeamento de *Y* para *Y*. Nesta equação, k é o número de amostras históricas necessárias para prever a resposta. A Equação 9 mostra o mesmo conceito.

$$f : Y^{-1}, Y^{-2}, ..., y^{-k} \rightarrow Y^{+1}$$

Equação 8 – Modelo Preditor Temporal

$$y^{+1} = f(y^{-1}, y^{-2}, ..., y^{-k})$$

Equação 9 – Modelo de um Sistema Preditor Temporal

Existem duas razões básicas para tratar alguns sistemas como uma caixa preta (Makridakis et al., 1997). Primeiro, o sistema pode ser pouco conhecido ou as dependências funcionais podem ser difíceis de serem estabelecidas. Segundo, é possível que uma determinada análise esteja focada tão somente na predição, não havendo interesse nas possibilidades de um modelo explanatório, tais como a criação de cenários e simulações.

[4] É oportuno destacar a existência de modelos que não são puramente explanatórios ou temporais, mas incorporam caracterísiticas de ambos. Estes modelos são alimentados por séries históricas de diferentes variáveis explanatórias (eventualmente, uma dessas entradas poderia ser uma série histórica da própria variável-resposta). Poderiam, por esta razão, ser classificados como *modelos híbridos*, mas a bibliografia consultada ao longo desta pesquisa reconhecia apenas os tipos puros (explanatórios ou temporais).

A Figura 11 mostra um esquema teórico, mostrando um preditor temporal baseado na máquina de aprendizado de Vapnik (Figura 9). Como a predição está num instante de tempo posterior ao presente, a exatidão da resposta não pode ser avaliada senão a posteriori ou durante o aprendizado, conforme sugerem a Figura 10 e a Equação 7.

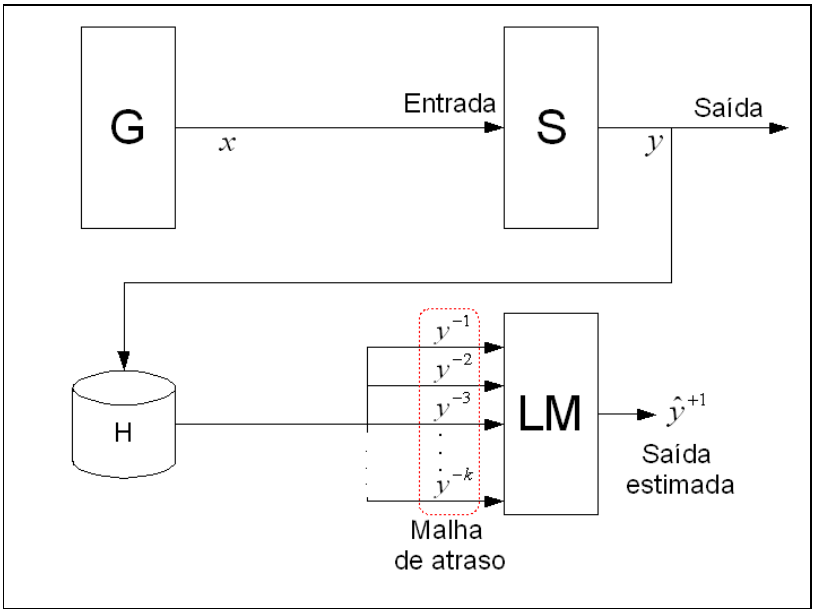

Figura 11 – Modelo Preditor Temporal (malha de atraso)

Em relação à Figura 9, a Figura 11 apresenta uma diferença significativa: a inclusão de uma base de dados históricos *H*, que armazena os dados necessários para a definição dos conjuntos de treinamento *S*.

3.4 Teoria de Aprendizado Estatístico e Minimização de Risco

Os algoritmos de aprendizado empírico dependem de uma figura de mérito para avaliar a qualidade da modelagem[5]. Na abordagem não-paramétrica, empregada pelos modelos conexionistas, o aprendizado é iterativo e o ajuste do modelo deve ser verificado a cada iteração. Em geral, o aprendizado é considerado completo (ou bem sucedido) quando a figura de mérito atinge um determinado valor.

[5] A figura de mérito é referida na literatura de estatística paramétrica como *medida de adequação do modelo*.

Uma figura de mérito popular é o erro médio quadrático (RMSE, *Root Mean Squared Error*), que transforma o aprendizado em um problema de minimização convexa (caso do *backpropagation*, utilizado pelas redes neuronais *perceptron*). Então, em princípio, o aprendizado consiste na busca de um conjunto de parâmetros tal que o RMSE seja mínimo. Todavia, freqüentemente a minimização conduz a uma anomalia conhecida como *perda de generalização* (Haykin, 1998).

Em seus trabalhos, Vladimir Vapnik (Vapnik, 1998; Vapnik 1999) contempla essa questão, propondo que o aprendizado empírico não fosse determinado somente pelo desempenho (RMSE), mas sim por um *compromisso* entre o desempenho e a complexidade dos modelos gerados. Desta forma, a questão da perda de generalização pode ser contornada sem prejuízo do desempenho obtido.

3.4.1 Considerações iniciais

Considerando a Equação 4 e a Equação 5, que definem a dependência funcional $X \rightarrow Y$ e o conjunto de treinamento $S = X \times Y$, assume-se que a dependência funcional assinalada pela equação 2 não é determinística, no sentido de que cada valor de x não determina um valor de y mas, sim, uma distribuição de probabilidade $P(X,Y)$, mostrada na Equação 10.

$$P(X,Y) = P(X) \cdot P(Y \mid X)$$

Equação 10 – Distribuição de Probabilidade Conjunta de X e Y

A Equação 10 mostra $P(X,Y)$ como uma distribuição de probabilidade desconhecida que gera o conjunto S independente e idênticamente (as chamadas condições i.i.d., *indenpendent and identically distributed*) (Cristiani, 2001). Portanto, $X \rightarrow Y$ é uma dependência probabilística descrita por $P=P(X,Y)$.

O aprendizado estatístico consiste em definir um estimador f (definido na Equação 4) que seja capaz de predizer o valor de y para qualquer $x \in X$ a partir de S. Esta estratégia é chamada de *modelagem empírica*, razão pela qual f é também chamado de *modelo empírico* do sistema (Vapnik, 1999).

A Equação 11 define f como membro de uma grande classe ou família de funções F, também denominado de espaço de hipóteses. Portanto, a Equação 4 admite muitas

soluções distintas, restando definir uma figura de mérito para selecionar qual delas é mais adequada.

$$F = \{f_i\}_{i=1}^{C}$$

Equação 11 – Classe de Funções (estimadores)

A Teoria do Aprendizado Estatístico define essa figura de mérito em termos do *risco funcional*, mostrado na Equação 12 (Vapnik, 1998). O risco funcional mede a quantidade média de erro (risco) associado a um estimador *f*, criando assim uma figura de mérito para selecionar o melhor estimador de *F*, denominado de f_0 (função alvo), como sendo aquele que apresenta o menor risco.

$$R = R(f_i)$$

Equação 12 – Risco Funcional de f_i

O risco funcional é definido em função dos erros de estimação de f_i. A Figura 9 e a Figura 11 destacam este fato, ao utilizar notações diferentes para as saídas real (*y*, a saída do sistema) e estimada ($\hat{y}$). Nesta seção, ambas as saídas serão denotadas respectivamente como *y* e *f(x)*. A função de perda (*loss function*) (Vapnik, 1998; Cristiani, 2001; Hastie et al., 2003) mostrada na Equação 13 define uma penalização para os erros de estimação. A literatura propõe uma série de implementações para esta função, sendo a mais popular mostrada na Equação 14, chamada de *perda do erro quadrático* (*squared-error loss function*) (Hastie et al., 2003).

$$L = L(y, f(x))$$

Equação 13 – Função de Perda

$$L = L(y, f(x)) = [y - f(x)]^2$$

Equação 14 – Função de Perda do Erro Quadrático

A partir da noção de perda, é possível definir o risco funcional como o erro de estimação esperado (EPE, *expected prediction error*) na forma da esperança matemática

mostrada na Equação 15 (Vapnik, 1998; Hastie et al., 2003). Como assinalado, o cálculo do cálculo da esperança (*E*) exige que a função de perda deva ser integrável para qualquer $f_i \in F$.

$$R(f_i) \equiv E[L(Y, f_i(X))] = \int_{X,Y} L(y, f_i(x)) \cdot P(x, y) dx dy$$

Equação 15 – Esperança Matemática Definida como *Risco Funcional Esperado*

A Equação 16 mostra a esperança da Equação 15 condicionada a *X*. O condicionamento é obtido fatorando a distribuição conjunta $P(X,Y) = P(Y \mid X) \cdot P(X)$, onde $P(Y \mid X) = P(Y, X)/P(X)$, e dividindo a integral bivariada de acordo (Hastie et al., 2003).

$$R(f_i) \equiv E_X E_{Y|X} [L(y, f_i(x)) \mid X]$$

Equação 16 – Função de Risco Condicionada a X

Com base na definição da Equação 15, o aprendizado estatístico pode ser formalmente definido como a busca por uma função $f_i \in F$ tal que $R(f_i)$ seja mínimo. A literatura destaca que os problemas de estatística básica relacionados à estimação de funções (tais como mínimos quadrados ou vizinho-mais-próximo) podem descritos como formas de minimizar o risco funcional a partir de uma coleção empírica de dados (Vapnik, 1998; Hastie et al., 2003). A Equação 17 mostra a função de risco condicionada da Equação 16 minimizada ponto a ponto, cuja solução é mostrada na Equação 18 (Hastie et al., 2003).

$$f(x) = \arg\min_c E_{Y|X} \left((Y - c)^2 \mid X = x \right)$$

Equação 17 – Minimização do Risco Ponto a Ponto

$$f_0(x) = E(Y \mid X = x)$$

Equação 18 – Esperança Condicionada

No jargão da estatística, a Equação 18 é a função de regressão que modela a dependência $X \rightarrow Y$ (Hastie et al., 2003). Embora se admita a existência de um estimador f_0 tal que $R(f_0)$ é mínimo, ele não pode ser determinado porque $P(X,Y)$ (que define o

risco na Equação 15) não é conhecido. Tudo que se conhece sobre o sistema é uma amostra de dados – o conjunto *S*. Para superar esta limitação, é utilizado um princípio de indução para aproximar *R* a partir desta amostra. Na modelagem empírica, o processo de indução deve ser tal que produza um estimador capaz de estimar respostas para dados que foram omitidos em *S* (poder de generalização).

3.4.2 Princípios de indução

A Teoria do Aprendizado Estatístico, tal como concebida por Vapnik (1999), define dois princípios de indução: Princípio da Indução da Minimização do Risco Empírico (ERM, *Empirical Risk Minimization*) e o Princípio da Indução da Minimização do Risco Estrutural (SRM, *Structural Risk Minimization*). O ERM é a forma clássica de minimização de risco empregado pelos primeiros algoritmos de aprendizado (modelos conexionistas), sendo de larga aceitação no meio científico. O SRM é uma contribuição original dos trabalhos de Vapnik sobre o SVM (*Support Vector Machine*), desenvolvido a partir da década de 90 (Vapnik, 1998; Vapnik, 1999; Cristiani, 2001; Schlkopf et al., 2001).

Na modelagem empírica, a quantidade e a qualidade das amostras de *S* determinam o desempenho do estimador obtido. Dado a sua natureza observacional, os dados são finitos e amostrados; por isso, problemas de amostragem são comuns, gerando dados não balanceados (Pindyck et al., 2004; Berry et al., 2004). Devido à alta dimensionalidade do problema, *S* normalmente é uma região esparsa do espaço das entradas. Conseqüentemente, a estimação será quase sempre um problema mal posto (*ill posed*), no sentido que diversas soluções são admitidas (Herbrich, 2001).

As abordagens conexionistas estão sujeitas a problemas de generalização, freqüentemente produzindo estimadores superajustados que apresentam um desempenho ruim após o aprendizado (Haykin, 1998). Isto é uma conseqüência direta dos algoritmos de otimização empregados para seleção de parâmetros[6], além das métricas estatísticas

[6] *Seleção de parâmetros* é um termo consagrado na literatura de redes neuronais, não possuindo relação com o conceito de *estimação de parâmetros* empregada pela estatística paramétrica. Refere-se ao processo de determinar quais variáveis de entradas (ou *parâmetros de entrada*) são relevantes para modelar o sistema em questão

usadas para selecionar a melhor solução. A qualidade de uma solução é dada pela minimização do risco funcional que, no caso das abordagens conexionistas, é obtido pela aplicação do ERM.

O ERM minimiza o erro das estimações obtidas durante o treinamento. Ao final do aprendizado, assume-se que o desempenho obtido com amostras novas será similar àquele obtido com o conjunto de treinamento, o que nem sempre é verdadeiro. A Equação 15 mostra a definição do risco esperado em termos da perda (função de perda) e da distribuição *P(X,Y)*. Como a distribuição não é conhecida, o risco só pode ser calculado por meio de uma aproximação baseada na informação disponível: o conjunto *S* e as propriedades de *F*; ou seja, o conjunto que contém a família de estimadores *f*.

A Equação 19 mostra uma aproximação para o risco esperado: o *risco empírico*. Tal risco é a média aritmética simples da função de perda para cada amostra do conjunto *S*. O Teorema Chave da Teoria do Aprendizado (Vapnik, 1998; Vapnik, 1999) demonstra que o risco empírico converge assintoticamente para o risco empírico quando o tamanho da amostra é muito grande; ou seja, quando $L \to \infty$. Portanto, uma alta cardinalidade de *S* garante uma solução empírica satisfatória com $R_{emp}(f_i) \to R(f_i)$.

$$R_{emp}(f_i) = \frac{1}{L} \cdot \sum_{l=1}^{L} L(f_i(x_l), y_i)$$

Equação 19 – Risco Empírico

Por outro lado, quando a amostra não é suficientemente representativa, a diferença entre R_{emp} e $R(f_i)$ pode diferir por uma larga margem, acarretando problemas de superajuste. Em termos práticos, isso significa que a minimização do risco empírico da Equação 19 não assegura sempre um erro de generalização pequeno.

A Figura 12 mostra como a minimização empírica (ERM) pode conduzir a anomalias, especialmente quando *S* é pequeno ou a sua amostragem é desbalanceada. Nesta figura, é mostrado como a minimização do erro durante o aprendizado não conduz necessariamente a uma diminuição do erro de generalização.

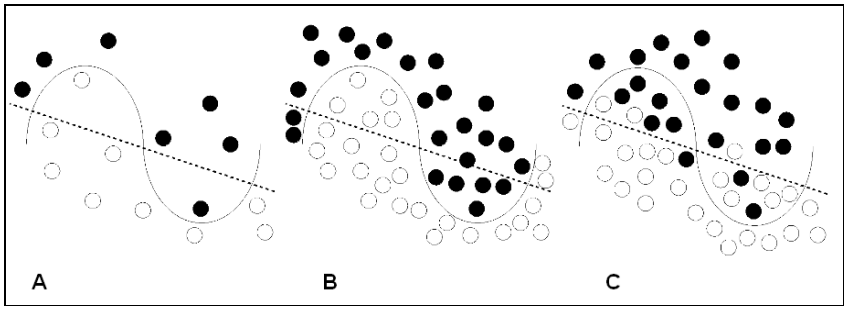

Figura 12 – O Dilema do Superajuste

Na Figura 12, existem duas classes de amostras identificadas por círculos de cores diferentes num espaço de características $\Re^2$. O problema consiste em determinar uma função discriminante tal que seja possível determinar a classe de uma amostra a partir das suas características. Portanto, o estimador pretendido é um classificador capaz de identificar as amostras com uma resposta binária, digamos $y \in \{+1,-1\}$. Em (a), um conjunto muito pequeno de dados é utilizado para o treinamento. Com base neste conjunto, são definidas duas hipóteses de classificação: uma discriminante complexa (a curva contínua) e uma simples (a reta tracejada). Claramente, o desempenho da discriminante contínua é superior à da tracejada em (a), pois a primeira hipótese classifica as amostras corretamente enquanto a segunda comete dois erros de classificação.

A Figura 12(a) pode ser explicada em termos do erro de treinamento: a hipótese contínua é complexa, mas apresenta um erro de treinamento inferior à da hipótese tracejada. Entretanto, este resultado aparentemente bom pode estar associado a um erro de generalização grande, o que seria percebido com amostras maiores.

A Figura 12(b) e a Figura 12(c) analisam o desempenho das mesmas hipóteses na presença de mais dados, omitidos no treinamento. Se a hipótese contínua for correta, corroborando o desempenho do treinamento, a hipótese tracejada seria subajustada, classificando mal as amostras na situação (b). Se a hipótese tracejada estiver correta, a hipótese contínua seria superajustada, também classificando mal as amostras em (c) a despeito do bom desempenho no treinamento mostrado em (a). Nesta última hipótese, o pequeno erro cometido pela hipótese tracejada no treinamento (a) acaba conduzindo a um melhor desempenho numa amostra maior de dados.

O superajuste reside neste dilema: um pequeno erro de treinamento pode acabar degradando o desempenho do estimador, diminuindo seu poder de generalização. Portanto, pode ser melhor admitir um erro de treinamento maior para assegurar um grau de generalização maior. A questão que surge é: até que ponto o erro de treinamento é válido como critério para seleção dos estimadores gerados?

Na prática, essa questão pode ser respondida através de um artifício denominado *validação cruzada* (Haykin, 1998; Duda et al., 2000). A validação cruzada propõe que um conjunto especial de dados, denominado conjunto de validação, seja utilizado para avaliar o estimador durante o processo de aprendizado.

O conjunto de treinamento é utilizado para gerar iterativamente novos estimadores que minimizam cada vez mais o erro de treinamento. O conjunto de validação, por sua vez, é utilizado para verificar o erro de generalização a cada iteração, sem interferir diretamente no aprendizado. Como estes conjuntos são disjuntos (não possuem elementos em comum), a utilização desta abordagem pode auxiliar a evitar estimadores superajustados.

A Figura 13 mostra como a validação cruzada pode ser empregada para minimizar o erro de generalização em uma situação de aprendizado típica. No gráfico, a ordenada mostra o desempenho dos estimadores gerados como o erro empírico, percebido como RMSE (*Root Mean Squared Error*)[7]. A abscissa representa as iterações utilizadas para gerar cada estimador (épocas), sendo uma medida discreta do tempo. Assim, para a época 1 se tem um par estimador/erro, para a época 2 outro e assim sucessivamente.

[7] O RMSE é a forma mais popular de se medir o desempenho de um estimador (Haykin, 1998; Duda et al., 2000). Tecnicamente, o RMSE é uma implementação do erro empírico (Equação 19) com a perda quadrática (Equação 14), sendo portanto definido como $RMSE(f_i) = \frac{1}{I} \cdot \sum_{i=1}^{I} L(f_i(x_i), y_i) = \sqrt{\frac{1}{I} \cdot \sum_{i=1}^{I} (f_i(x_i) - y_i)^2}$

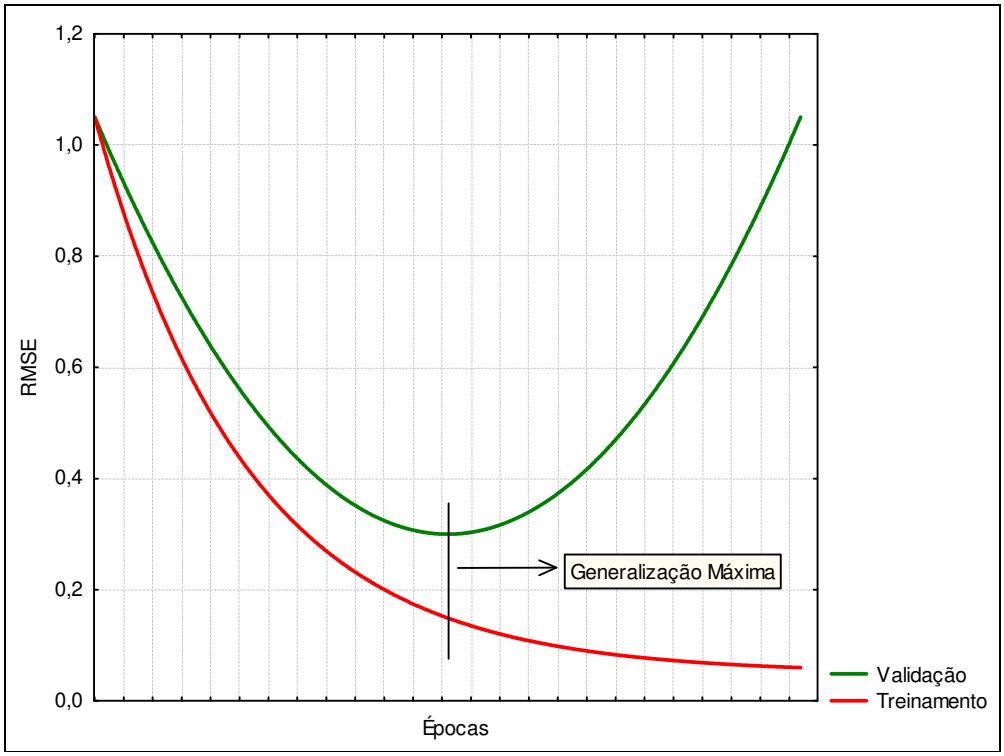

Figura 13 – Validação Cruzada e Generalização

A curva vermelha da Figura 13 mede o RMSE do treinamento de um estimador numa aplicação de regressão de função. Como pode ser observado, esse erro tende a diminuir à medida que o aprendizado progride e gera estimadores melhor adaptados ao conjunto de treinamento. Tipicamente, o estimador associado a cada época[8] possui um desempenho superior (RMSE menor) do que o estimador da época anterior, razão pela qual a curva é decrescente. Para aplicações de regressão de funções, a curva será sempre assintótica, possivelmente tangenciando o nível de erro zero quando o tempo (época) for relativamente grande.

A curva verde mostra o erro de validação, que é o desempenho obtido pelo estimador (RMSE) no processamento do conjunto de validação. A geração dos estimadores não é influenciada pelo conjunto de validação, o qual tem por única finalidade verificar o poder de generalização dos estimadores depois que eles são criados.

[8] *Época* é empregada na literatura de redes neuronais e de Aprendizado Estatístico como sendo um passo (ou iteração) do processo iterativo de ajuste dos parâmetros ajustáveis de um estimador.

O erro de generalização é sempre maior do que o obtido com o conjunto de treinamento, embora as duas curvas tendam a decrescer juntas até um ponto denominado *generalização máxima* (assinalado na Figura 13). A partir deste ponto, o modelo começa a perder o poder de generalização e, apesar do erro de treinamento continuar a cair, o erro de generalização começa a aumentar. Neste caso, o comportamento do risco esperado aproxima-se do erro de generalização, que pode ser grande mesmo quando o erro de treinamento se aproxima de zero. Por essa razão, a literatura recomenda que o aprendizado seja interrompido exatamente no ponto de generalização máxima (Haykin, 1998; Duda et al., 2000; Hastie et al., 2003).

A validação cruzada contorna a questão da generalização do ERM, permitindo uma conciliação entre desempenho e generalização. Entretanto, é necessário haver amostras suficientes para compor independentemente os conjuntos de treinamento e de validação. Além disso, a seleção do melhor estimador continua sendo um problema mal posto: pode haver muitas, possivelmente infinitas, funções f_i que minimizam o erro empírico.

Por outro lado, a Figura 12 mostra que um erro de treinamento pequeno (erro empírico) pode estar associado a uma hipótese complexa que apresenta um alto erro de generalização. É precisamente este ponto que o SRM explora para minimizar o risco esperado: restringir a complexidade dos estimadores gerados no aprendizado. De certa forma, o SRM é uma aplicação da navalha de Occam[9], ao considerar que um estimador simples que explique a maioria dos dados de S com um erro pequeno é melhor do que um estimador complexo com desempenho igual ou mesmo superior (Figura 12). A idéia é definir um estimador pertencente a uma classe de função (Equação 11), cuja complexidade seja a menor possível.

Existe um paradoxo potencial no risco estrutural: a minimização de complexidade pode aumentar o risco empírico, o que conduziria a um erro esperado maior. Então, a busca de soluções no aprendizado não pode se limitar a poucas classes de funções restri-

[9] A navalha (ou crivo) de Occam é oriunda dos trabalhos de Guilherme de Occam, frade franciscano do séc. XIV que é considerado um dos precursores do pensamento crítico científico. O original em latim, *pluralitas non est ponenda sine neccesitate*, é traduzido livremente como *dada duas teorias que explicam igualmente os mesmos fatos, a mais simples deve ser a correta.* No contexto da TAE, se uma função

tas, de pequena complexidade, mas sim considerar muitas classes diferentes de funções. A melhor solução possível é um compromisso entre o erro empírico e a complexidade da solução utilizada para minimizar este erro (Vapnik, 1998). Este compromisso é a base qualitativa do SRM.

A Equação 20 mostra a complexidade de um estimador, conhecido como *fator de regularização*, *termo de confiança* ou *termo de capacidade* (Schlkopf et al., 2001), denotado por Φ. Φ é uma função da capacidade *h*, que mede a complexidade da classe *F*, do número *L* de amostras presentes no conjunto de treinamento e de uma probabilidade mínima η. Intuitivamente, para uma mesma probabilidade η, se a capacidade *h* é grande e o número de amostras *L* é pequeno, então a distância entre os erros empírico e esperado tende a ser grande também, o que conduz a erros de generalização grandes.

$$\Phi = \Phi\left(\sqrt{\frac{h}{L}}, \eta\right)$$

Equação 20 – O Fator de Regularização Φ (limite probabilístico do risco empírico)

O fator de regularização é utilizado durante o aprendizado para controlar a complexidade do estimador usado para explicar *S* (Schlkopf et al., 2001), razão pela qual também é chamado de *termo de complexidade*. A regularização fornece limites probabilísticos para a distância entre os riscos esperado e empírico de qualquer função, incluindo o minimizador do risco empírico em um espaço de funções que pode ser usado para controlar o superajuste (Vapnik, 1999).

A relação entre o risco esperado, o risco empírico e o termo de regularização é mostrada por meio da desigualdade da Equação 21.

$$R(f_i) \leq R_{emp}(f_i) + \Phi\left(\sqrt{\frac{h}{L}}, \eta\right)$$

Equação 21 – O Risco Esperado Relacionado ao Risco Empírico e ao Fator de Regularização Φ

---

simples explica uma amostra empírica de dados, não há necessidade de buscar outra mais complexa. Por esta razão, o crivo também é conhecido como *Princípio da Economia*.

A TAE define muitas formas de medir a capacidade *h* de uma classe de funções *F*, sendo que a mais comum é a chamada *dimensão VC* (dimensão de Vapnik e Chervonenkis) (Vapnik, 1998; Herbrich, 2001). O termo "dimensão" neste contexto tem uma conotação diferente da usual, ao designar uma medida da capacidade de um algoritmo de classificação binária. Essa medida é dada pelo maior conjunto de pontos *p* que tal algoritmo pode distinguir dicotomicamente em todas as 2^p maneiras possíveis (Vapnik, 1998; Hastie et al., 2003).

Em geral, os discriminantes lineares possuem a capacidade *h* igual ao número de parâmetros livres mais um; ou seja, quanto maior o grau de liberdade, maior a capacidade do discriminante e, conseqüentemente, maior a complexidade envolvida (Herbrich, 2001; Hastie et al., 2003). Outros tipos de discriminante, como a sinusóide, podem ter capacidade infinita (Vapnik, 1999; Hastie et al., 2003). Para aplicações de regressão, onde $f(x) \in \Re$, a capacidade de *f(x)* é medida indiretamente por meio de uma função indicadora, nos termos da Equação 22 (Hastie et al., 2003). Esta função indicadora possui como parâmetros *f(x)* e β, uma constante que assumes valores na mesma faixa que *f(x)*, retornando dois valores possíveis (classificação binária). Neste caso, a capacidade h de *f(x)* será adequadamente aproximada por *I* (Hastie et al., 2003).

$$I = I(f(x) - \beta)$$

Equação 22 – A Complexidade de *f(x)* Medida Indiretamente por Meio de uma Função Indicadora

A Equação 21 determina um risco esperado mínimo $R(f_i)$ através de um erro de treinamento $R_{emp}(f_i)$ pequeno e de um estimador tão simples quanto possível (fator de regularização Φ pequeno). Essa abordagem admite duas situações extremas:

a) Uma classe de função F_k produz um fator de regularização baixo, mas um alto erro no treinamento é observado.

b) Uma classe de função $F_{k'}$ produz um erro de treinamento baixo, mas o fator de regularização obtido é grande.

A melhor classe de função está entre os dois extremos mostrados, conforme mostra a Figura 14, dado que o interesse é obter uma função (estimador) que explique os

dados do treinamento suficientemente bem e que mantenha o risco esperado (erro de generalização) baixo.

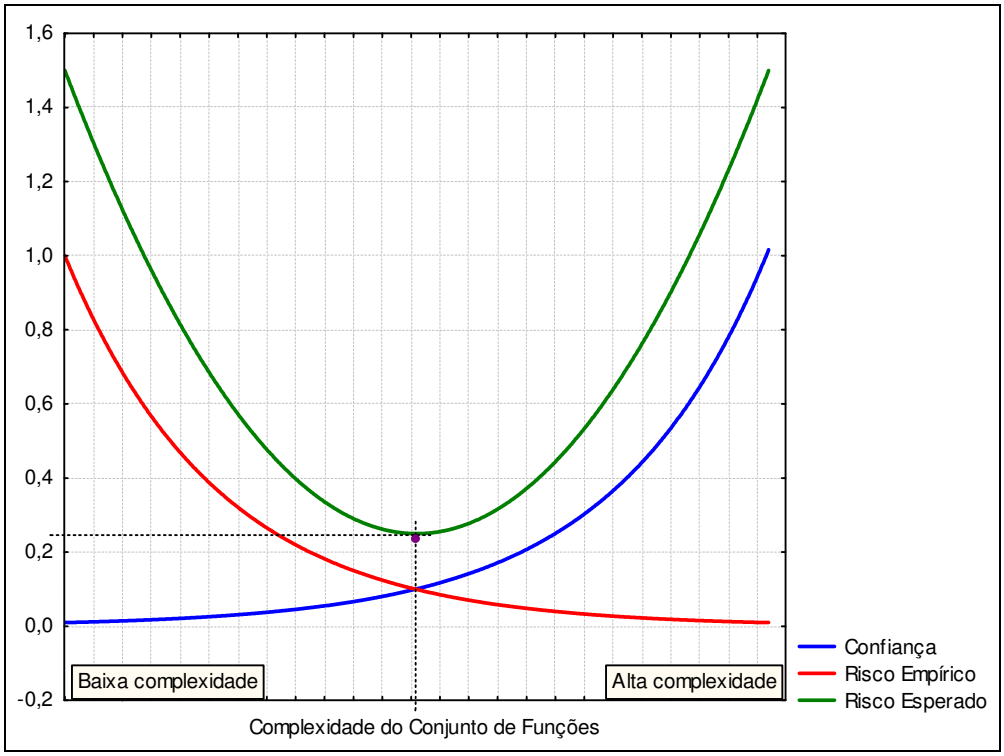

Figura 14 – Limites Probabilísticos entre os Riscos Empírico e Esperado

A Figura 14 mostra graficamente a desigualdade da Equação 21, onde são traçados o erro empírico (erro de treinamento, em vermelho), o limite superior do termo de complexidade (em azul) e o risco esperado correspondente (em verde). A análise da figura mostra que uma complexidade alta assegura um erro empírico baixo, mas o limite superior da confiança se torna pior. O ponto mínimo do erro esperado (assinalado na figura) é atingido para a capacidade h de uma determinada classe F. Portanto, o aprendizado consiste num ponto de equilíbrio entre o erro empírico e a complexidade Φ.

A inspeção da Figura 13 e da Figura 14 mostra que o risco esperado e o erro de generalização estão fortemente correlacionados. Na primeira figura, o número de épocas (abscissa do gráfico) torna os modelos cada vez mais especializados, e a partir do ponto de máxima generalização ocorre perda de generalização. Na segunda, os riscos esperado e empírico caem à medida que a capacidade h cresce; quando o limite probabilístico começa a se tornar muito alto, o risco esperado começa a subir enquanto o erro empírico

continua caindo. Ambas as figuras sugerem que é necessário balancear a capacidade *h* e o erro empírico, sem o quê o risco esperado (ou o erro de generalização) pode comprometer o desempenho do estimador gerado. Portanto, há indícios que, durante o processo iterativo de aprendizado, os estimadores gerados vão se tornando cada vez mais complexos. Como conseqüência, fica estabelecido que uma das formas de controlar a complexidade dos estimadores gerados é restringir o número de iterações do processo de aprendizado.

A Figura 15 mostra o processo de aprendizado por meio de uma analogia gráfica. Nesta analogia, os membros de uma classe de funções são representados num *espaço de hipóteses*, que é um subconjunto de todos os estimadores possíveis (*espaço alvo*).

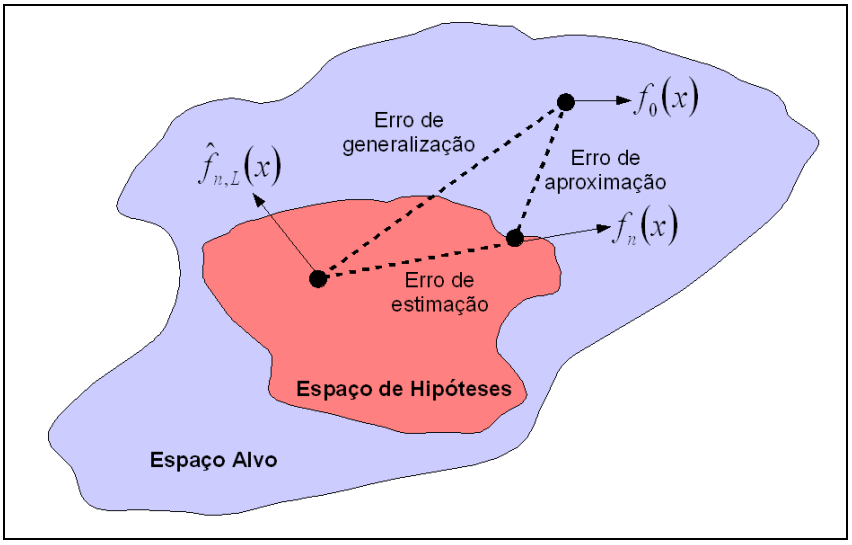

Figura 15 – Erros no Aprendizado

Na Figura 15, erro de generalização é representado como a soma vetorial do erro de estimação e do erro de aproximação. O erro de estimação é conseqüência do procedimento de aprendizado, que em função de uma série de limitações (como a escolha de parâmetros ou um número de amostras insuficiente) acaba produzindo um modelo sub-ótimo ($\hat{f}_{n,L}$) dentre aqueles disponíveis em *F* – na figura, o melhor modelo disponível é f_n. O erro de aproximação é uma conseqüência do espaço de hipóteses ser menor do que o espaço alvo – assim, se a função alvo f_0 estiver fora do espaço de hipótese, o melhor modelo que pode ser construído será necessariamente sub-ótimo.

A classe de funções *F* pode ser demasiadamente grande, razão pela qual o aprendizado normalmente considera espaços de hipóteses menores *H*. O SRM propõe que seja empregada uma seqüência aninhada de espaços de hipóteses $H_1 \subset H_2 \subset H_3 \subset ... \subset H_M$, onde cada espaço tenha uma capacidade finita h_m maior do que todos os conjuntos nele contido; ou seja, $h_1 < h_2 < h_3 < ... < h_M$. Por exemplo, H_m poderia ser o conjunto de polinômios de grau m.

4 Máquinas de Vetores de Suporte (*Support Vector Machines*)

4.1 Introdução

O SVM (*Support Vector Machine*, Máquina de Vetores de Suporte) é um conjunto de algoritmos supervisionados usados para classificação e regressão, constituindo uma opção atraente para modelagem empírica de dados. O método explora a idéia de regularização, mencionada no Capítulo 3. Com isso, limita-se a complexidade da função gerada no aprendizado enquanto que o erro de generalização é mantido sobre controle. Outra noção importante empregada pelo SVM é o chamado *princípio da margem* que não somente permite evitar o superajuste como contorna a *maldição da dimensionalidade* (seção 3.2). Com efeito, o SVM não depende explicitamente da dimensionalidade do espaço de entrada, mas somente do produto interno de vetores, que é um valor real. Isso permite a construção de hiperplanos de classificação em espaços de alta dimensionalidade, até mesmo em espaços de Hilbert com dimensionalidade infinita (Vapnik, 1998).

Uma das aplicações mais comuns da teoria do aprendizado é a classificação que, assim como a estimação, é uma forma de aprendizado empírico (indução estatística). A idéia é construir um estimador que, a partir de um conjunto de treinamento, seja capaz de classificar amostras desconhecidas com um erro mínimo (capacidade de generalização). A Equação 23 mostra um conjunto de treinamento para um problema de classificação binária, onde L é o total de amostras e cada classe é identificada por um rótulo numérico matematicamente conveniente (+1 ou -1).

$$S = \{(x_1, y_1), (x_2, y_2), ..., (x_L, y_L)\} \in X \times Y \quad X \in \Re^N, Y = \{\pm 1\}$$

Equação 23 – Conjunto de Treinamento

Não existe nenhuma premissa a respeito de *X*, exceto que se trata de um conjunto de dados onde certos vetores multidimensionais *x* são associados à *y*. A forma como

esses vetores de *X* são determinados a partir de uma situação no mundo real (extração de características) é completamente abstraída.

Para que a classificação seja possível, algum tipo de similaridade deve ser definido tanto em *X* como em *Y*, de tal sorte que *similaridades na entrada impliquem em similaridades na saída*. Para aplicações de classificação, a similaridade na saída é obtida diretamente porque os valores envolvidos são discretos e facilmente comparáveis: as saídas podem ser iguais ou diferentes. Em contra partida, a forma de similaridade na entrada é conceitualmente mais complexa de definir, constituindo um dos principais objetivos do aprendizado (Schlkopf et al., 2001).

A Equação 24 mostra uma função de similaridade *k*, que associa um número real – o grau de similaridade – a dois vetores de *X*. Na literatura de aprendizado estatístico, esta função é chamada de *kernel* (núcleo) (Schlkopf et al., 2001).

$$k : X \times X \rightarrow \Re \mid (x, x') \mapsto k(x, x')$$

Equação 24 – Métrica de Similaridade em X

Uma forma bem conhecida de similaridade é o *produto interno*, que pode ser definido de diversas formas (Steinbruch, 1987 et al.; Schlkopf et al., 2001). O *produto interno canônico* é uma dessas formas, mostrada na Equação 25. Esta métrica, que só é precisa quando os vetores comparados são normalizados, é numericamente igual ao co-seno do ângulo observado entre esses vetores. Considerando que os vetores são normalizados, a Figura 16 mostra que o produto interno canônico (dado pelo co-seno de α) é uma medida indireta da distância euclidiana entre eles, dada por $\|x_1 - x_2\|$ (destacada na Figura 16 como retas tracejadas).

$$\langle x_1, x_2 \rangle = \sum_{i=1}^{N} x_{1_i} \cdot x_{2_i}$$

Equação 25 – Produto Interno Canônico

De fato, se x_1 e x_2 são similares em um determinado espaço de características, espera-se que eles estejam topologicamente próximos. Isso diminui o ângulo α e torna o

co-seno deste ângulo mais próximo de 1 (similaridade máxima). Caso contrário, maior é a distância entre os vetores, o ângulo observado é maior e o co-seno deste ângulo aproxima-se de 0 (similaridade mínima). Também é possível que a distância observada seja tal que o ângulo formado é obtuso, hipótese na qual o co-seno se aproximaria de -1 (dissimilaridade máxima).

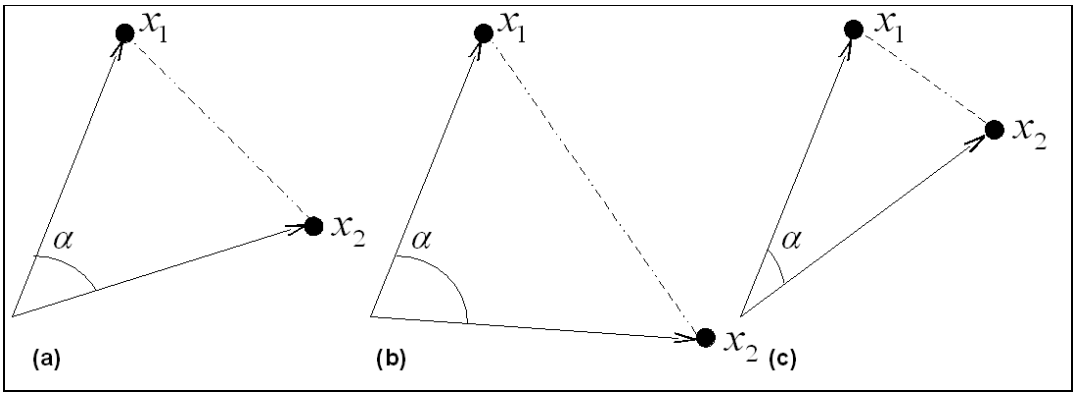

Figura 16 – O Produto Interno Canônico como Métrica de Similaridade

É importante notar que esta aplicação do produto interno é análoga à correlação estatística, que também mede a semelhança entre vetores (Steinbruch, 1987 et al.; Oppenheim et al., 1999; Montgomery et al., 2001).

O produto interno, em sua conotação mais ampla, não é definido para todo espaço geométrico (Schlkopf et al., 2001). Portanto, sua utilização é limitada a determinadas classes de espaços, denominados *espaços euclidianos* (Steinbruch, 1987 et al.) ou *espaços de Hilbert* (Vapnik, 1999; Schlkopf et al., 2001). Dado que a natureza do espaço de entrada pode ser muito diversificada, talvez seja necessário convertê-lo mediante um mapeamento apropriado (*extração de características*), tal como mostrado na Equação 26.

$$\Phi : X \rightarrow H_c$$
$$x \mapsto \vec{x} = \Phi(x)$$

Equação 26 – Mapeando o Espaço de Entrada X para o Espaço de Características H_c

Algumas técnicas de classificação ou regressão comumente utilizadas, como redes RBF (*Radial Basis Function Network*) ou MLP (*Multilayer Perceptron*), utilizam o mapeamento implicitamente, ao modificar o espaço de entrada original através de funções de base radial ou de camadas ocultas. A idéia do mapeamento Φ é representar os padrões de entradas como vetores num espaço de características H_c. Com isso, uma métrica de similaridade tal como o produto interno canônico pode ser aplicado, o que não é necessariamente possível no espaço original. Na Equação 26, essa idéia é reforçada pela notação empregada: enquanto os padrões de entrada são representados como *x*, os vetores de características correspondentes em *H* são denotados por $\vec{x}$.

A Equação 27 salienta que o uso do mapeamento permite calcular a similaridade de dois padrões de entrada (x_1 e x_2) mesmo num espaço onde o produto interno não está definido.

$$k(x_1, x_2) = \langle \vec{x}_1, \vec{x}_2 \rangle = \langle \Phi(x_1), \Phi(x_2) \rangle$$

Equação 27 – Similaridade no Espaço de Características

Em muitas situações, o mapeamento é imprescindível ao SVM, como nos casos em que os padrões não são linearmente separáveis. Tipicamente, o mapeamento é um operador não-linear que representa o espaço original em uma dimensão muito mais alta. Quando a dimensionalidade for suficientemente alta, as amostras transformadas poderão ser discriminadas por um hiperplano (Duda et al., 2000); ou seja, por um discriminante linear. Por esta razão, H_c é algumas vezes denominado de *espaço de linearização* (Schlkopf et al., 2001).

A Figura 17 mostra uma situação hipotética, onde um aumento de dimensionalidade torna a função de decisão mais simples. No gráfico da Figura 17a, os padrões não podem ser discriminados por uma função linear, ao contrário do que acontece com o espaço transformado de maior dimensionalidade mostrado no gráfico da Figura 17b.

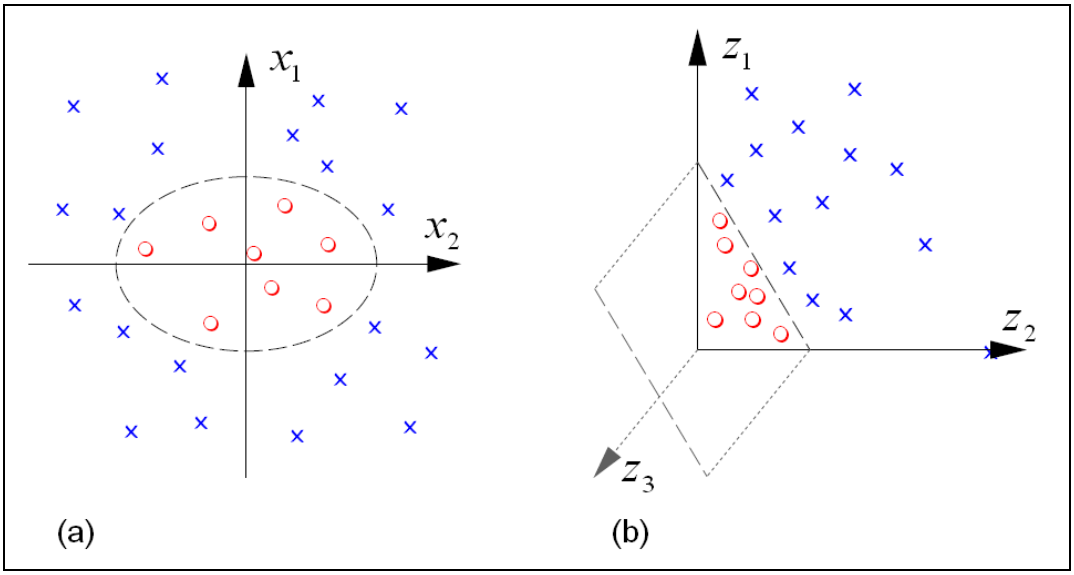

Figura 17 – Mapeamento do espaço $X \subset \Re^2$ para o espaço $Z \subset \Re^3$

Neste ponto, um paradoxo parece se estabelecer: por um lado, o mapeamento para um espaço de dimensionalidade maior possibilita a discriminação linear entre os padrões. Mas, em contrapartida, o princípio da maldição da dimensionalidade parece restringir a aplicabilidade deste tipo de mapeamento. De acordo com a maldição da dimensionalidade, a quantidade de padrões necessária para amostrar adequadamente o espaço de entrada é, em tese, uma função exponencial da dimensionalidade (Haykin, 1998; Duda et al., 2000). Em contrapartida, o mapeamento pode tornar o aprendizado muito mais simples, restringindo a capacidade (complexidade) do discriminante. Isso pode ser percebido pela Figura 17, onde a função de decisão a é consideravelmente mais complexa que em b, muito embora o primeiro espaço esteja no $\Re^2$ e o segundo, no $\Re^3$. É neste ponto que o SVM se distingue fundamentalmente de outras técnicas de classificação, como os discriminantes de Fischer: o aprendizado está muito mais fortemente ligado ao pré-processamento (o mapeamento) do que propriamente ao algoritmo de classificação (Duda et al., 2000).

Independentemente de o mapeamento ser *necessário*, ele pode ser *útil* para representar convenientemente os padrões de entrada. Como os padrões de entrada são representados como vetores no espaço de características, muitos conceitos úteis de álgebra linear e geometria analítica podem ser aplicados à teoria do aprendizado. Mesmo nos

casos onde existe um produto interno definido no espaço de entrada, o mapeamento pode ser útil se permitir o emprego de outras formas de similaridade mais adequadas ou mesmo o uso de outros algoritmos de aprendizado (Schlkopf et al., 2001).

4.2 Contextualização

Sem prejuízo da generalização, considere-se um problema de classificação binária, onde o objetivo é construir um discriminante que distinga duas classes de amostras (identificadas pelos rótulos numéricos $+1$ e -1) num determinado espaço de características. A Figura 18 mostra o cenário descrito, representando as duas classes como objetos diferentes num espaço de características $\Re^2$. A função de decisão (ou discriminante) pode ter diversas formas, como as linhas mostradas nesta figura.

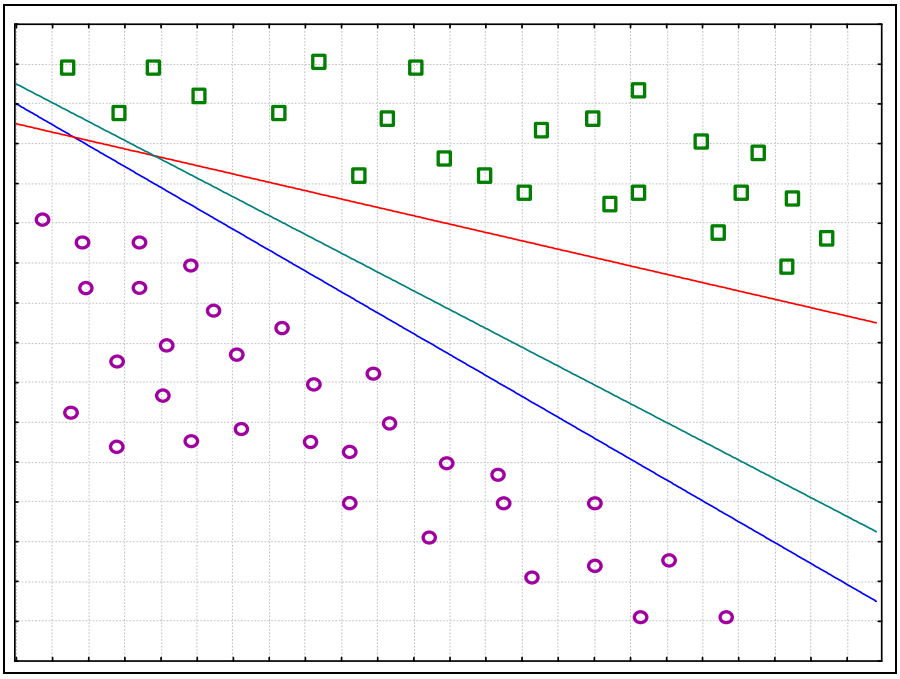

Figura 18 – Classificação Binária

O desempenho dos discriminantes da Figura 18 é satisfatório para todos os casos mostrados. Não obstante, não é possível determinar *a priori* qual desses discriminantes deve funcionar melhor com amostras desconhecidas, a menos que uma métrica específi-

ca seja concebida para isso. O conceito de vetor de suporte se aplica em situações como essa, determinando critérios para a escolha de um discriminante ótimo cujo desempenho seja probabilisticamente assegurado também com amostras desconhecidas. O SVM estabelece uma noção intuitiva: o melhor discriminante é aquele que maximiza a distância entre ele e os pontos mais próximos de cada classe. Esse classificador é único, sendo denominado de *hiperplano separador ótimo* (Vapnik, 1998; Schlkopf et al., 2001).

Existem abordagens paramétricas que apresentam soluções correlatas, tais como a análise de discriminantes (Duda et al., 2000; Rencher, 2002; Hastie et al., 2003). Entretanto, a abordagem paramétrica é limitada em algumas aplicações do mundo real devido a três fatores: maldição da dimensionalidade, incompatibilidade da função de distribuição real com as funções de distribuição estatísticas clássicas e restrições à aplicação do método da máxima verossimilhança para estimação de densidade (Vapnik, 1998). Essas limitações só podem ser contornadas quando a cardinalidade (número de elementos) do conjunto de amostras é muito grande em comparação com a sua dimensionalidade. Isso pode ser percebido mediante a inspeção da Figura 19, que mostra duas classes bivariadas normais delimitadas por elipses. Se estas duas classes forem multivariadas normais e possuírem matrizes de covariâncias semelhantes, é possível demonstrar que a separação proporcionada por qualquer discriminante pode ser expressa em uma nova dimensão (Rencher, 2002), tal como mostrado nessa figura.

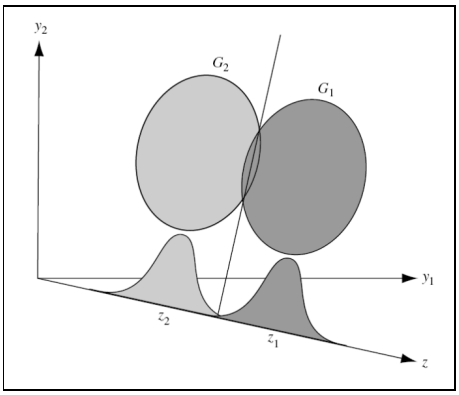

Figura 19 – Análise de Discriminante para Amostras de Duas Categorias

Notar que o espaço de características da Figura 19 é bidimensional, sendo constituído por y_1 e y_2. A dimensão z é uma representação geométrica alternativa à função

discriminante. A linha que une os pontos de intersecção das duas elipses pode ser projetada sobre a reta *z*, identificando assim o ponto de máxima separação entre as classes e o melhor discriminante (ponto de sobreposição mínima) (Rencher, 2002). Com efeito, as duas gaussianas traçadas sobre *z* estão bem espaçadas, mais do que seria possível com qualquer outro discriminante.

Não obstante a elegância e a simplicidade da solução paramétrica mostrada na Figura 19, é preciso destacar a forte influência que a amostragem exerce nesta abordagem. Em termos práticos, a forma como a amostragem é feita pode deslocar os centróides de cada elipse ou modificar o contorno de cada uma delas, o que fatalmente conduziria a conclusões deturpadas, prejudicando o cálculo do melhor discriminante. Além disso, existe uma premissa não declarada quando da utilização desses métodos: a invariância no tempo, que determina a constância dos parâmetros do modelo (média e variância, no caso da distribuição normal). Neste trabalho, ambos os quesitos (amostragem e invariância) não são completamente atendidos, o que justifica a escolha da minimização de risco estrutural com SVM em detrimento dos métodos paramétricos ou mesmo dos modelos conexionistas (minimização de risco empírico).

4.3 Algoritmo de Classificação Binário

A Equação 28 mostra os centróides das duas classes (rotuladas como $+1$ e -1) envolvidas num problema de classificação binária. O centróide de cada classe é um vetor médio ($\vec{c}_+$ e $\vec{c}_-$), computado como a média aritmética dos vetores de cada classe. Nessa equação, m_+ é a quantidade de amostras com rótulo $+1$, enquanto que m_- é a quantidade de amostras com rótulo -1.

$$\vec{c}_+ = \frac{1}{m_+} \sum_{\{i | y_i = +1\}} \vec{x}_i \quad \vec{c}_- = \frac{1}{m_-} \sum_{\{i | y_i = -1\}} \vec{x}_i$$

Equação 28 – Centróides das Classes (classificação binária)

A partir dos centróides, é possível aplicar a regra do vizinho mais próximo (*nearest neighbor*) (Duda et al., 2000) para identificar amostras desconhecidas (omitidas ou

não disponíveis durante o aprendizado). A Figura 20 mostra os centróides de alguns padrões representados num espaço de características plano ($\Re^2$).

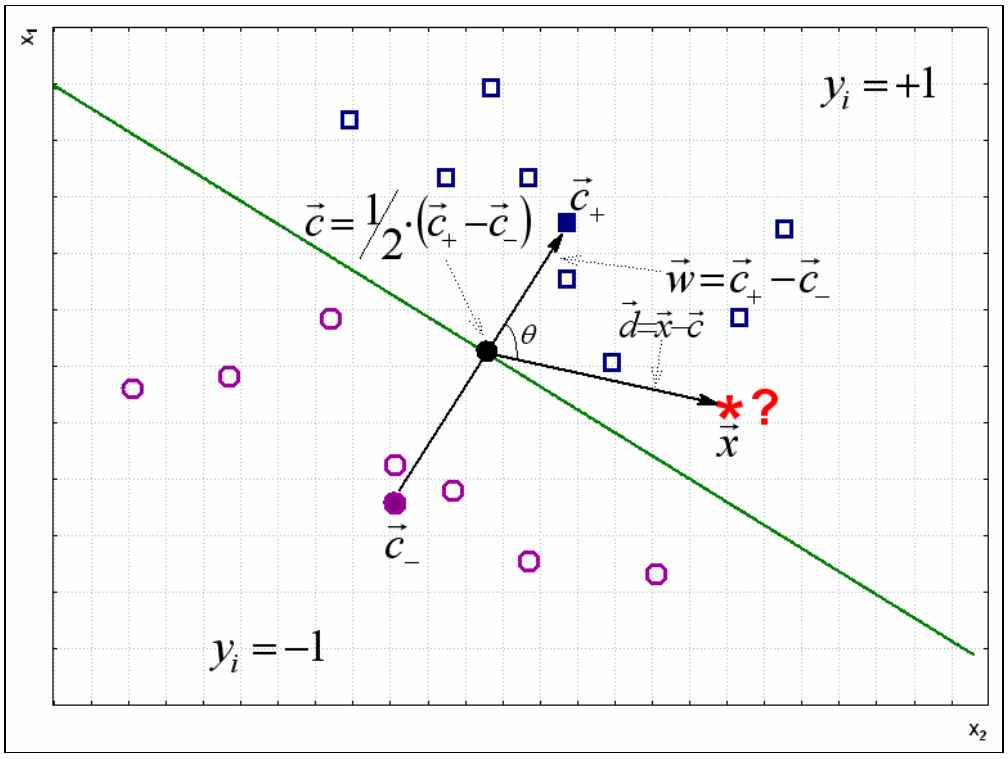

Figura 20 – Classificação Binária num Espaço de Características

Aplicando noções de álgebra e geometria analítica aos conceitos desenvolvidos, é possível definir um algoritmo de classificação ao mesmo tempo simples e eficiente. Considere-se o cenário descrito pela Figura 20, onde o objetivo é classificar uma amostra desconhecida $\vec{x}$ (mostrada como um asterisco vermelho) a partir das demais amostras. Seja $\vec{c}$ o vetor média dos centróides, $\vec{w}$ o vetor diferença dos centróides e $\vec{d} = \vec{x} - \vec{c}$ o vetor que une $\vec{c}$ à $\vec{x}$. A classificação de $\vec{x}$ é dada pela sua afinidade com um dos centróides, que pode ser medida pelo ângulo θ entre $\vec{w}$ e $\vec{d}$. A regra de decisão correspondente à Figura 20 é mostrada na Tabela 1 (Schlkopf et al., 2001).

Tabela 1 – Regras de Decisão para a Classificação Binária

Regra de decisão	Classe
$\theta < \pi/2$	+1
$\theta > \pi/2$	-1
$\theta = \pi/2$	Indefinido

A Equação 29 mostra essa regra de decisão expressada com o uso da função *de-grau* (sgn), cujo comportamento é definido na Equação 30 e mostrado graficamente na Figura 21. Como pode ser constatado, existe um ponto de descontinuidade no gráfico onde a função não é definida. Este ponto de descontinuidade é chamado de *limiar* que, no argumento da Equação 30, será nulo.

$$\begin{aligned} y &= \operatorname{sgn}\left(\left\langle (\vec{x}-\vec{c}), \vec{w} \right\rangle\right) \\ &= \operatorname{sgn}\left(\left\langle \vec{x} - (\vec{c}_+ - \vec{c}_-)/2, (\vec{c}_+ - \vec{c}_-) \right\rangle\right) \\ &= \operatorname{sgn}\left(\left\langle \vec{x}, \vec{c}_+ \right\rangle - \left\langle \vec{x}, \vec{c}_- \right\rangle + b\right) \end{aligned}$$

Equação 29 – Função de Decisão Empregando o Degrau

$$\operatorname{sgn}(a) = \begin{cases} +1, a > 0 \\ -1, a < 0 \end{cases}$$

Equação 30 – Definição da Função Degrau

A função degrau (também função *sinal, limiar* ou *treshold*) mapeia qualquer número real para $\{+1,-1\}$, não definindo nenhum outro valor para o seu contradomínio. Mediante a manipulação algébrica mostrada na Equação 29, essa propriedade permite implementar a regra de decisão da Figura 20, razão pela qual no aprendizado estatístico o degrau é chamado de *função de decisão*.

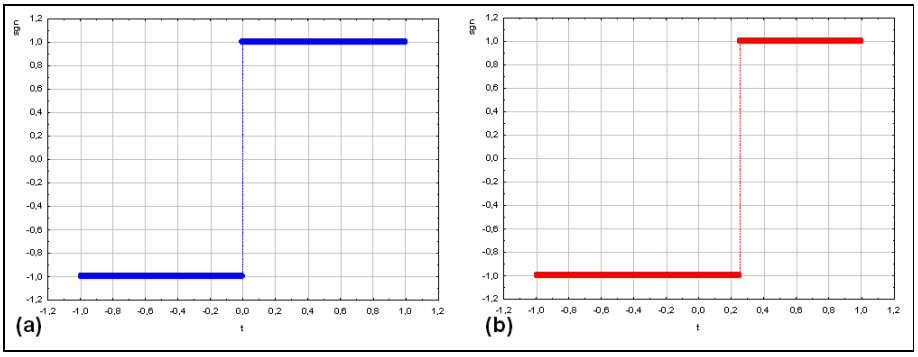

Figura 21 – A Função Sinal (sgn)

A Equação 29 define um desvio (*offset* ou *bias*) aplicado ao argumento da função degrau. Geometricamente, o desvio equivale a deslocar o limiar do degrau de 0 (Figura 21a) para b (Figura 21b). Sua aplicação é necessária quando os centróides das classes não são simétricos em relação à origem, exigindo uma compensação na função de decisão. A Equação 31 mostra que b é calculado a partir dos centróides, tendendo a diminuir quando seus módulos forem parecidos. Se este parâmetro fosse simplesmente omitido, o hiperplano é forçado a cruzar a origem dos eixos, restringindo a solução (Schlkopf et al., 2001).

$$b = \frac{1}{2} \cdot \left(\|\vec{c}_-\|^2 - \|\vec{c}_+\|^2 \right)$$

Equação 31 – Valor de Bias

4.4 Hiperplanos de classificação e o Hiperplano ótimo

É possível e conveniente fazer uso da função de decisão justamente no limiar mostrado na Equação 29 e na Equação 30. O limiar representa o limite de decisão entre as classes – de fato, a resposta do degrau no limiar é indefinida – determinando um hiperplano perpendicular ao vetor $\vec{w}$. Este hiperplano é formado pelos pontos cujas respostas ao degrau somadas ao efeito de b são nulas. Em particular na Figura 20, o espaço de características está no $\Re^2$, de tal sorte que o hiperplano formado é uma reta (traçada em verde).

A Equação 32 pode ser obtida a partir da Equação 29 e mostra como o hiperplano de classificação pode ser descrito pelo limiar de decisão (Vapnik, 1999; Schlkopf et al., 2001). Dependendo de $\vec{w}$, essa equação pode definir uma ampla família de hiperplanos aptos a discriminar corretamente os padrões de entrada. Essa família é linear em seus parâmetros, o que torna possível determinar a sua complexidade (capacidade).

$$\langle \vec{w}, \vec{x} \rangle + b = 0$$

Equação 32 – Hiperplano de Classificação

Como $\vec{w}$ ajusta a família de hiperplanos definida na Equação 32, ele é chamado de vetor de *ponderação*, de *pesos* ou de *parâmetros* (Haykin, 1998; Duda et al., 2000). A propriedade mostrada na Equação 33 ressalta que o *limite de decisão* representado pelo hiperplano de classificação não contém nenhuma amostra conhecida. Como conseqüência, toda amostra existente é obrigatoriamente classificada, não existindo ambigüidades (Schlkopf et al., 2001).

$$y_i \cdot \left(\langle \vec{w}, \vec{x}_i \rangle + b \right) > 0 \quad \forall i = 1,2,...,m$$

Equação 33 – Propriedade do Vetor de Parâmetros $\vec{w}$

Uma ambigüidade de classificação seria notada como um ponto contido no hiperplano ou então muito próximo dele, fazendo com que a resposta de uma função de decisão seja indefinida. No caso da função degrau (Equação 30), isso equivale à resposta do degrau justamente no limiar ou muito próximo dele.

A Equação 34 mostra a função de decisão correspondente à família de hiperplanos definida na Equação 32.

$$y = \operatorname{sgn}\left(\langle \vec{w}, \vec{x} \rangle + b \right)$$

Equação 34 – Função de Decisão

A solução da Equação 32 é indeterminada; ou seja, admite várias soluções. Sem prejuízo da generalidade, é imprescindível restringi-la de alguma forma. Então, por convenção, exige-se que o hiperplano de classificação seja *cartesiano*. Isso significa re-

escalonar o vetor $\vec{w}$ e a constante b de tal forma que a Equação 33 se restringe à forma da Equação 35.

$$y_i \cdot \left(\left\langle \vec{w}, \vec{x}_i \right\rangle + b\right) \geq 1 \quad \forall i = 1,2,...,m$$

Equação 35 – Forma Canônica do Hiperplano de Classificação

Independentemente da forma canônica, a constante b deve satisfazer as duas desigualdades mostradas na Equação 36 (Vapnik, 1998) para que a função de decisão da Equação 34 efetue a classificação de forma consistente. Como pode ser notado, a resposta não é definida no limiar b.

$$(a) \quad \left\langle \vec{w}, \vec{x} \right\rangle > -b \rightarrow y = +1$$
$$(b) \quad \left\langle \vec{w}, \vec{x} \right\rangle < -b \rightarrow y = -1$$

Equação 36 – Desigualdades de Decisão

A literatura (Vapnik, 1999; Duda et al., 2000; Schlkopf et al., 2001) propõe diversos algoritmos para problemas de classificação onde os padrões de entrada podem ser separados por hiperplanos; ou seja, situações onde os padrões de entrada são *linearmente separáveis* (Haykin, 1998; Duda et al., 2000). Entre esses algoritmos, existe um denominado *panorama generalizado* (*generalized portrait*), baseado em duas noções fundamentais (Schlkopf et al., 2001).

A primeira noção é a de *hiperplano ótimo*, caracterizado pela *máxima margem* de separação possível entre as diferentes classes. A Equação 37 mostra como o hiperplano ótimo pode ser interpretado analiticamente: a *menor* distância ($\min\left(\left\|\vec{x} - \vec{x}_i\right\|\right)$) entre uma amostra de cada classe (x_i) e o hiperplano ($\vec{x} \mid \left\langle \vec{w}, \vec{x} \right\rangle + b = 0$) deve ser a *maior* possível ($\max\left\{\min\left(\left\|\vec{x} - \vec{x}_i\right\|\right)\right\}$). Se a forma canônica expressa pela Equação 35 não for observada, a Equação 37 torna-se indeterminada; ou seja, não admite solução única (Vapnik, 1998).

$$\max\left\{\min\left(\left\|\vec{x} - \vec{x}_i\right\|\right)\right\} \middle| \vec{x} \in H_c, \left\langle \vec{w}, \vec{x} \right\rangle + b = 0, i = 1,2,...,m$$

Equação 37 – Condições para a Obtenção do Hiperplano Ótimo

A segunda noção é que a capacidade de separação dos hiperplanos diminui à medida que a largura da margem cresce, o que explica a boa generalização do hiperplano ótimo (Schlkopf et al., 2001). A largura da margem é um conceito análogo ao intervalo de confiança da estatística descritiva, mensurando qualitativamente a precisão do classificador gerado: em geral, quanto maior a margem, maior a esperança que a resposta fornecida pelo classificador está correta (*princípio da margem*) (Herbrich, 2001).

4.5 Cálculo do hiperplano ótimo

A resolução da Equação 37 requer algumas considerações sobre a forma como a regra de decisão é implementada. A regra mostrada Equação 29 é similar à da Equação 34, mas a forma como os hiperplanos são obtidos é diferente. No primeiro caso, a Figura 20 mostra que o vetor normal ao hiperplano ($\vec{w}$) é calculado mediante um procedimento simples: a diferença entre os centróides ($\vec{w} = \vec{c}_+ - \vec{c}_-$).

No segundo caso, a estratégia empregada é mais elaborada e $\vec{w}$ é obtido resolvendo a Equação 38 para o máximo valor da margem ρ (ou seja, $\max(\rho)$), mostrada na Equação 39 (Vapnik, 1998). É possível demonstrar que a margem $\rho(\vec{w})$ é máxima dentro da região $\|\vec{w}\| \leq 1$, sendo atingida no limite $\|\vec{w}\| = 1$. Este ponto máximo é uma decorrência da continuidade de ρ na área delimitada por $\|\vec{w}\| \leq 1$ (Vapnik, 1998).

$$c_{-1}(\vec{w}) = \min\langle \vec{w}, \vec{x}_i \rangle \quad y_i = -1$$
$$c_{+1}(\vec{w}) = \max\langle \vec{w}, \vec{x}_i \rangle \quad y_i = +1$$

Equação 38 – Parâmetros para Determinar o Hiperplano Ótimo

$$\rho(\vec{w}) = \frac{c_{+1}(\vec{w}) - c_{+1}(\vec{w})}{2} \qquad \|\vec{w}\| = 1$$

Equação 39 – Margem de Separação entre as Classes

Considere-se um vetor $\vec{w}_0$ tal que $\rho(\vec{w}_0)$ é o maior valor possível para ρ que satisfaz as desigualdades da Equação 36 e o ponto médio c_0, definido na Equação 40. Então, $\vec{w}_0$ e c_0 determinam um hiperplano de classificação que assegura a máxima separação entre as classes: o hiperplano ótimo, mostrado na Equação 41 (Vapnik, 1998).

$$c_0 = \frac{c_{-1}(\vec{w}_0) + c_{+1}(\vec{w}_0)}{2}$$

Equação 40 – Ponto Médio da Margem de Separação

$$\langle \vec{w}_0, \vec{x} \rangle + c_0 = 0$$

Equação 41 – Equação do Hiperplano Ótimo

A forma ótima do hiperplano mostrada na Equação 41 é canônica, possuindo portanto as propriedades da Equação 42, que são uma decorrência da definição mostrada na Equação 35 (Schlkopf et al., 2001).

$$\begin{aligned} \langle \vec{w}_0, \vec{x}_i \rangle + c_0 \geq +1 \quad y_i = +1 \\ \langle \vec{w}_0, \vec{x}_i \rangle + c_0 \leq -1 \quad y_i = -1 \end{aligned}$$

Equação 42 – Restrições de Desigualdade da Forma Canônica

Considere-se a função mostrada na Equação 43, denominada de *função objetivo*. Dentro das restrições impostas pela forma canônica definida na Equação 35, é possível demonstrar o valor de $\vec{w}$ que satisfaz a Equação 43 está relacionado à $\vec{w}_0$ pela relação mostrada na Equação 44 (Vapnik, 1998).

$$\min \tau(\vec{w}) = \frac{1}{2} \cdot \|\vec{w}\|^2 = \frac{1}{2} \cdot \langle \vec{w}, \vec{w} \rangle$$

Equação 43 – Função Objetivo para Determinar o Hiperplano Ótimo

$$\vec{w}_0 = \frac{\vec{w}}{\|\vec{w}\|}$$

Equação 44 – Relação entre $\vec{w}_0$ e $\vec{w} \mid \|\vec{w}\|^2 = \min\{\|\vec{w}\|^2\}, y_i \cdot \left(\langle \vec{w}, \vec{x}_i \rangle + b\right) \geq 1$

Juntas, as restrições da Equação 42 e a função objetivo da Equação 43 formam o chamado *problema de otimização restrita* (*constrained optimization problem*) (Schlkopf et al., 2001). Determinar o hiperplano ótimo consiste em solucionar este problema; ou seja, atingir a função objetivo observando as restrições de desigualdade. Matematicamente, devido à forma quadrática da Equação 43, o problema da otimização restrita é considerado uma questão de *otimização quadrática* (Vapnik, 1998). Uma das formas de tratar a otimização quadrática é mapeando $\vec{w}$ e b para um espaço dual – mais especificamente, o espaço dos multiplicadores de Lagrange (Vapnik, 1998; Vapnik, 1999; Schlkopf et al., 2001; Hastie et al., 2003).

O Lagrangiano é uma solução matemática padrão para problemas de otimização tal como o descrito. A Equação 45 mostra o Lagrangiano que deve ser *minimizado* com respeito às variáveis fundamentais $\vec{w}$ e b ao mesmo tempo em que deve ser *maximizado* para as variáveis duais não negativas α_i, o que equivale a procurar por um ponto de cela na superfície definida por esta função (Vapnik, 1998; Schlkopf et al., 2001). A forma como o Lagrangiano pode ser resolvido requer a aplicação de princípios de cálculo para analisar os pontos extremos de sua superfície, o que foge ao escopo deste trabalho. Contudo, informações detalhadas sobre a resolução do Lagrangiano podem ser obtidas em (Vapnik, 1998; Vapnik, 1999; Cristiani, 2001; Schlkopf et al., 2001, Hastie et al., 2003).

$$L(\vec{w}, b, \alpha) = \frac{1}{2} \cdot \|\vec{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i \left(\langle \vec{x}_i, \vec{w} \rangle + b \right) - 1 \right)$$

Equação 45 – Lagrangiano

O Lagrangiano da Equação 45 pode ser aplicado à função de decisão mostrada na Equação 34, produzindo a função de decisão visualizada na Equação 46, onde $\vec{x}$ é a amostra a ser classificada e $\vec{x}_i$ é a i-ésima amostra do conjunto de treinamento.

$$y = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \cdot \alpha_i \langle \vec{x}, \vec{x}_i \rangle + b \right)$$

Equação 46 – Função de Decisão Empregando os Multiplicadores de Lagrange

A Equação 44 mostra que o vetor $\vec{w}_0$ é normalizado; isto é, possui norma unitária. A normalização é imprescindível porque a norma dos vetores obtidos durante o processo de aprendizado pode variar, o que tornaria inconsistente a comparação entre as diversas soluções possíveis (Herbrich, 2001).

A Equação 47 mostra a largura da margem obtida com o hiperplano ótimo (Vapnik, 1998).

$$\rho(\vec{w}_0) = \sup_{\vec{w}_0} 1/2 \cdot \left(\min_{y_i=+1} \langle \vec{x}_i, \vec{w}_0 \rangle - \max_{y_i=-1} \langle \vec{x}_i, \vec{w}_0 \rangle \right) = \frac{1}{\|\vec{w}\|} \qquad \vec{w}_0 = \frac{\vec{w}}{\|\vec{w}\|}$$

Equação 47 – Margem Obtida com o Hiperplano Ótimo

A Figura 22 mostra o conceito de hiperplano ótimo graficamente. As amostras mais próximas do hiperplano (circuladas em vermelho) determinam a largura da margem, sendo chamadas de *vetores de suporte* (Vapnik, 1998; Duda et al., 2000; Schlkopf et al., 2001). É de grande interesse notar que o hiperplano de separação, ótimo ou não, pode ser determinado tão somente pelos vetores de suporte, sendo independente das demais amostras (Schlkopf et al., 2001).

Se o problema é linearmente separável, existe um vetor de pesos $\vec{w}$ e um limiar b tal que $y_i \cdot (\langle \vec{w}, \vec{x}_i \rangle + b) > 0$, onde x_i é a i-ésima amostra do conjunto de treinamento. Redimensionando $\vec{w}$ e b de tal forma que $\|\langle \vec{w}, \vec{x} \rangle + b\| = 1$ é satisfeita para os pontos de cada classe que estão mais próximos do hiperplano (x_1 e x_2, realçados com círculos tracejados), diz-se que o hiperplano está na forma canônica (Equação 35). Neste caso, a distância entre os vetores de suporte, mostrada na Equação 48, é igual à $2/\|\vec{w}\|$ (Schl-

kopf et al., 2001). Como o hiperplano é linear, essa largura é mantida ao longo de todo o espaço de características.

$$\left.\begin{array}{l}\langle \vec{w}, \vec{x}_1 \rangle + b = -1 \\ \langle \vec{w}, \vec{x}_2 \rangle + b = +1\end{array}\right\} \langle \vec{w}, \vec{x}_2 - \vec{x}_1 \rangle = 2 \Rightarrow \left\langle \frac{\vec{w}}{\|\vec{w}\|}, \vec{x}_2 - \vec{x}_1 \right\rangle = \frac{2}{\|\vec{w}\|}$$

Equação 48 – Distância entre os Vetores de Suporte (largura da margem)

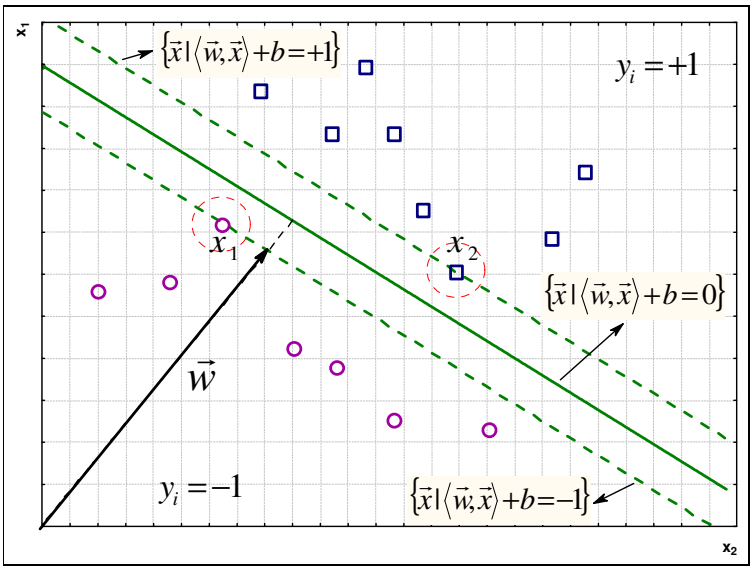

Figura 22 – Hiperplano Ótimo e Vetores de Suporte para uma Classificação Binária

4.6 Máquinas de Vetores de Suporte

Os conceitos de mapeamento e de similaridade no espaço de características definem um algoritmo de aprendizado especialmente poderoso: as máquinas de vetores de suporte. A idéia essencial deste algoritmo é bem ilustrada pelos gráficos da Figura 17, onde uma função de decisão complexa num espaço de baixa dimensionalidade (Figura 17a) é substituída por uma função de decisão linear num espaço de dimensionalidade maior (Figura 17b). Desta forma, conforme salientado na seção 1, se, por um lado, a complexidade aumenta devido ao aumento de dimensionalidade, por outro as regras de decisão tendem a se tornar mais simples.

Não obstante, a despeito do poder de generalização do hiperplano ótimo ser comprovado, determiná-lo é apenas uma possibilidade teórica (Vapnik, 1998; Vapnik, 1999). Com efeito, a dimensionalidade do espaço de características tende a ser muito mais alta do que a do espaço original (Duda et al., 2000) e, conseqüentemente, o problema pode se tornar inviável computacionalmente. Por exemplo, um discriminante polinomial de grau 4 ou 5 num espaço original $\Re^{200}$ exige um espaço de características de bilhões de dimensões para formar hiperplanos separadores (Vapnik, 1999). A existência de funções de decisão simples (hiperplanos) em espaços de características de alta dimensionalidade é uma certeza teórica; entretanto, determiná-los pode ser uma tarefa computacional intratável.

Entretanto, Vapnik e outros pesquisadores (Vapnik, 1998; Vapnik, 1999) determinaram que o mapeamento, embora realmente necessário para assegurar a existência de funções de decisão simples, não precisava ser considerado na forma explícita. De fato, a Equação 27 mostra que a similaridade no espaço de características pode ser determinada sem efetuar o mapeamento, através do *kernel* k. Assim, o *kernel* torna possível determinar o hiperplano sem que seja necessário efetuar o mapeamento explicitamente. Matematicamente, isso é assegurado pelo Teorema de Mercer ou condições de Mercer (Vapnik, 1998; Vapnik, 1999; Cristiani, 2001; 80).

A existência da expansão do *kernel* k mostrada na Equação 49, assegura a existência de um espaço de características H_c tal que k é o seu produto interno. Para que esta expansão seja possível, o teorema de Mercer associa toda função $g \in L2$ ao kernel k por meio de uma integral dupla, mostrada na Equação 50. Se a integração for positiva, a expansão é possível e k descreve um produto interno.

$$k(x_1, x_2) = \sum_{i=1}^{\infty} \alpha_i \Phi(x_1) \Phi(x_2) \qquad \alpha_k > 0$$

Equação 49 – Expansão do Kernel k

$$\iint k(x_1, x_2) g(x_1) g(x_2) d_{x1} d_{x2} \qquad \int g(a) d_a < \infty$$

Equação 50 – Condição de Mercer

É justamente a convolução do produto interno que permite construir funções de decisão não-lineares (alta complexidade) no espaço de entrada, as quais equivalem a funções de decisão lineares nos espaços de características de alta dimensionalidade criados por Φ (k é uma convolução do produto interno para este espaço de características) (Schlkopf et al., 2001).

Substituir $\langle \Phi(x_1), \Phi(x_2) \rangle = \langle \vec{x}_1, \vec{x}_2 \rangle$ por $k(x_1, x_2)$, tal como a Equação 24 sugere ser possível, é um artifício simples conhecido como *truque do kernel* (*kernel trick*) (Schlkopf et al., 2001; Herbrich, 2001). Tal artifício é suficientemente poderoso para estender a funcionalidade dos hiperplanos formados pelo panorama generalizado (discutido na seção 4.4) ao expressar a similaridade entre os vetores de H_c diretamente a partir de X. A forma da função de decisão é mostrada no termo à direita da Equação 51 (Schlkopf et al., 2001). Como se observa, esta expressão faz uso dos multiplicadores lagrangianos $\alpha_i \geq 0$, mas omite o mapeamento Φ.

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i \langle \Phi(x), \Phi(x_i) \rangle + b \right) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b \right)$$

Equação 51 – Decisão Implícita no Espaço de Características

Os algoritmos que definem funções de decisão com a forma do termo à direita da Equação 51 são chamadas de *máquinas de vetores de suporte*. Com efeito, o argumento da função de decisão ($\sum y_i \alpha_i k(x, x_i) + b$) somente precisa considerar os vetores de suporte, omitindo as demais amostras (Vapnik, 1999).

O termo *máquina de vetores de suporte* enfatiza a idéia básica de expansão dos vetores de suporte na forma do termo à direita da Equação 51. É por essa razão que no SVM a complexidade do aprendizado está relacionada ao número de vetores de suporte, e não à dimensionalidade do espaço de características (Vapnik, 1999). A Figura 23 mostra o diagrama esquemático de uma máquina de vetores de suporte, cuja arquitetura é semelhante a um MLP de duas camadas.

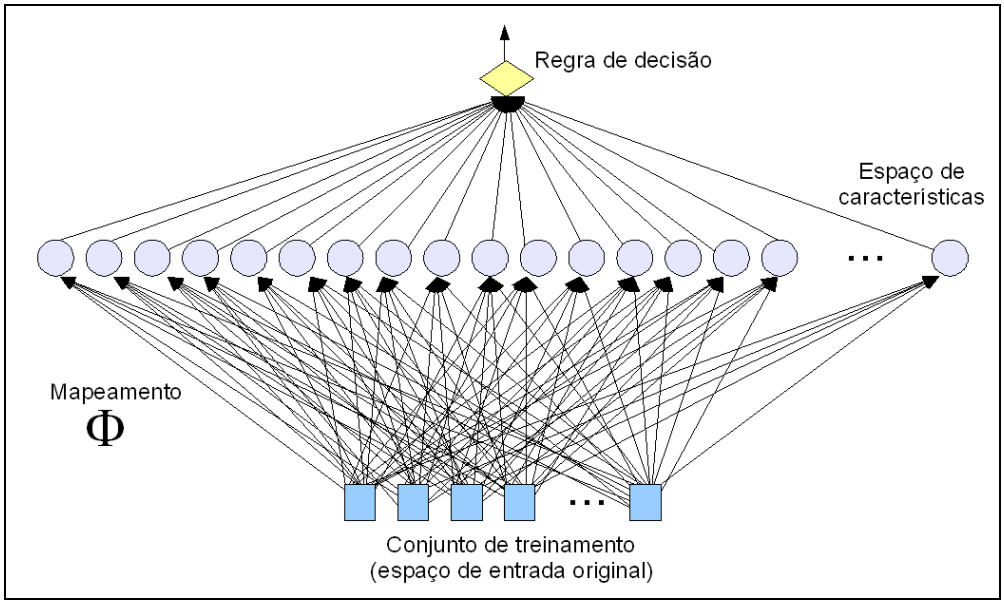

Figura 23 – Arquitetura de uma Máquina de Vetores de Suporte

A Figura 24 mostra um classificador obtido por meio de um kernel de base radial mostrada na Equação 52 (a chamada *base radial gaussiana*) (Schlkopf et al., 2001). O discriminante é mostrado como uma linha sólida verde, enquanto que as linhas tracejadas delimitam a margem, atendendo a restrição da forma canônica expressa na Equação 35. Isso significa que os vetores de suporte (amostras circuladas em vermelho) se encontram justamente sobre aquelas linhas, de maneira que, se i é o índice de qualquer uma desses vetores, $y_i \cdot \left(\langle \vec{w}, \vec{x}_i \rangle + b\right) \equiv 1$.

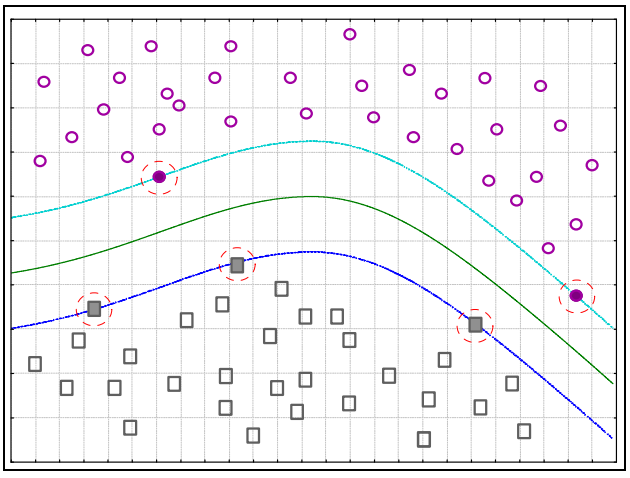

Figura 24 – Função de Decisão no Espaço de Entrada

$$k(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2} \right)$$

Equação 52 – Kernel de Função de Base Radial

Na Figura 24, quanto mais afastada as amostras se encontram do limite de decisão, maior é o valor de $\left\| \sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b \right\|$ (módulo do argumento da função de decisão) (Schlkopf et al., 2001).

4.7 Hiperplanos de Margens Suaves (*Soft Margin hyperplanes*)

Por razões práticas, pode ser necessário relaxar o critério da forma canônica mostrado na Equação 35, permitindo assim alguns erros de classificação durante o aprendizado. Até então, o objetivo era maximizar a largura da margem, assegurando assim um erro de generalização (esperado) muito baixo. Mediante a aplicação do truque do kernel, é sempre possível restringir a complexidade do classificador, tornando-o linear em algum espaço de características.

A noção de maximização da margem de classificação é um ponto de partida conceitual para a construção de algoritmos mais robustos adequados a situações reais (Cristiani, 2001). Por exemplo, sob condições não-ideais, é provável que a amostragem do conjunto de treinamento esteja contaminada por algum tipo de ruído, o que pode causar uma sobreposição entre as classes (Schlkopf et al., 2001). Em geral, isso *impede* a separação linear entre as classes *a menos* que sejam empregadas funções de kernel tão poderosas como a função radial definida na Equação 52. Nestas condições, o classificador pode se tornar super ajustado, produzindo uma larga discrepância entre o risco empírico e o risco real (Cristiani, 2001; Herbrich, 2001). Outra não idealidade que pode afetar a separação linear é a ocorrência de pontos discrepantes, que pode tornar a convergência extremamente lenta (Herbrich, 2001).

Essas deficiências do algoritmo podem ser superadas com o emprego de uma heurística denominada *SVM de margem suave* (*soft margin SVM*). Uma margem suave permite o relaxamento na forma canônica do discriminante (Equação 35), através da introdução de parâmetros livres, como mostrado na Equação 53 (Vapnik, 1998; Vapnik, 1999, Schlkopf et al., 2001).

$$y_i \cdot \left(\left\langle \vec{w}, \vec{x}_i \right\rangle + b\right) \geq 1 - \xi_i \quad \xi_i \geq 0 \quad \forall i \in \{1,2,...,m\}$$

Equação 53 – Relaxamento da Forma Canônica

De certa forma, ao admitir erros de classificação, o relaxamento da Equação 53 torna o classificador quase-ótimo. Entretanto, em muitas situações reais esta é a única maneira de manter o erro de generalização sob controle, o que torna esta solução efetivamente superior.

Para determinar o classificador, é necessário modificar a função objetivo original (Equação 43) para admitir uma taxa de erro, limitando-a de forma conveniente, tal como mostrado na Equação 54 (Schlkopf et al., 2001).

$$\min \tau\left(\vec{w}, \vec{\xi}\right) = \frac{1}{2} \cdot \left\|\vec{w}\right\|^2 + C \cdot \sum_{i=1}^{m} \vec{\xi}_i \qquad C > 0$$

Equação 54 – Função Objetivo para Determinar o Hiperplano Quase-Ótimo (condições relaxadas)

Com esta nova função objetivo, a determinação do classificador passou a depender do controle da complexidade (via $\left\|\vec{w}\right\|$) e do somatório $\sum_{i=1}^{m} \vec{\xi}_i$, onde C é uma constante que permite encontrar um ponto de equilíbrio entre a maximização da margem e a minimização do erro de treinamento. Como no caso do hiperplano ótimo, a função objetivo é atingida com a utilização de um kernel e de multiplicadores de Lagrange, os quais estão sujeitos à restrição da Equação 55.

$$0 \le \alpha_i \le C \qquad \sum_{i=1}^{m} \alpha_i y_i = 0 \quad \forall i \in \{1,2,...,m\}$$

Equação 55 – Restrições para o Lagrangiano

O Lagrangiano e outros aspectos matemáticos relacionados são omitidos por extrapolarem o escopo deste estudo; no entanto, sua resolução é uma extensão natural do Lagrangiano utilizado para determinar o hiperplano ótimo. Considerações mais aprofundadas são mostradas em (Vapnik, 1998; Vapnik, 1999, Schlkopf et al., 2001).

4.8 Regressão SVM *(SVR* – Support Vector Regression*)*

A análise de regressão distingue-se fundamentalmente da classificação por considerar muitos (possivelmente infinitos) valores numéricos contínuos como resposta (ou seja, $y_i \in \Re$), ao invés de um pequeno número de possibilidades categóricas (discretas). Por esta razão, a regressão pode ser considerada uma forma de generalização da classificação, onde a resposta esperada varia continuamente dentro de uma faixa determinada.

A regressão SVM é semelhante à determinação dos hiperplanos de margem suave, porém com a adoção de uma função de perda. Uma função de perda comum é a função ε-insensitiva de Vapnik, mostrada na Figura 25 (Vapnik, 1998; Vapnik, 1999; Schlkopf et al., 2001; Herbrich, 2001). Entretanto, outras funções de perda também são populares, tais como o erro quadrático e suas variações, a função de Huber, o Laplaciano ou as polinomiais (Vapnik, 1998; Vapnik, 1999, Schlkopf et al., 2001).

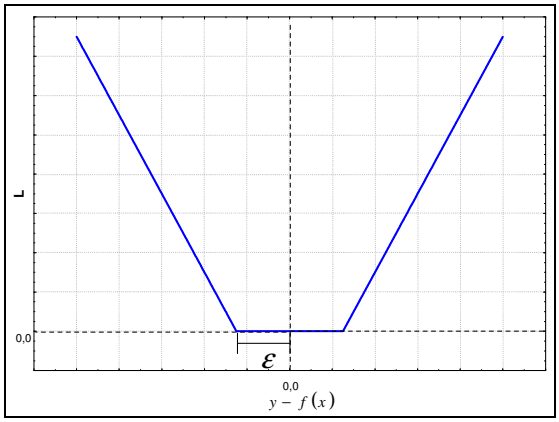

Figura 25 – Função de Perda ε-insensitiva de Vapnik

O gráfico da Figura 25 mostra a penalização ou perda (L) em função do erro observado ($y - f(x)$). Quanto maior o erro, maior a penalização e se não houver erro, não há penalização ($L = 0$). Deve-se notar que só há penalização na função ε-insensitiva quando o módulo do erro extrapola um valor arbitrário ε. Isso forma uma região de tolerância determinada por $(-\varepsilon, +\varepsilon)$ (região *insensitiva*), visualizada como um platô na parte inferior da curva mostrada na Figura 25. Fora da faixa de tolerância, todo erro é penalizado em proporção linear ao seu módulo.

A Equação 56 expressa algebricamente a função de perda da Figura 25.

$$c(x, y, f(x)) = |y - f(x)|_{\varepsilon} = \max\{0, |y - f(x)| - \varepsilon\}$$

Equação 56 – Função de Perda ε-insensitiva de Vapnik

A região de tolerância pode ser visualizada na regressão mostrada na Figura 26, onde a curva ajustada (em vermelho) é calculada levando em conta somente os erros cujos módulos extrapolam ε. Isso determina um tubo de raio ε ao redor da curva ajustada. Então, o melhor modelo é aquele cuja capacidade é relativamente baixa e o número de pontos fora do tubo é relativamente pequeno.

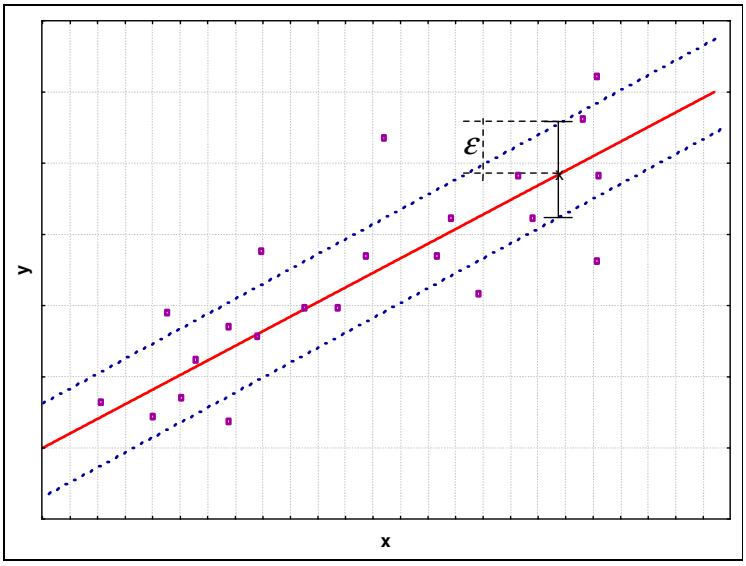

Figura 26 – Regressão SVM

Como no caso da classificação, a estimação de uma função linear na forma da Equação 57 envolve um fator de otimização quadrática, mostrada na Equação 58. Este fator é utilizado pela função objetivo mostrada na Equação 59.

$$f(\vec{x}) = \langle \vec{w}, \vec{x} \rangle + b$$

Equação 57 – Regressão Linear

$$\tfrac{1}{2} \cdot \|\vec{w}\|^2 + C \cdot \sum_{i=1}^{m} \|y_i - f(x_i)\|_\varepsilon$$

Equação 58 – Fator de Otimização Quadrática

$$\min\left(\tau\left(\vec{w}, \vec{\xi}^{(*)}\right)\right) = \min\left(\tfrac{1}{2} \cdot \|\vec{w}\|^2 + C \cdot \sum_{i=1}^{m} \left(\xi_i + \xi_i^*\right)\right)$$

Equação 59 – Função Objetivo para a Regressão Linear

A função objetivo da Equação 59 contém duas variáveis livres: ξ e ξ^*, denotadas coletivamente como $\vec{\xi}^{(*)}$. Elas são introduzidas para limitar o erro lateralmente; ou seja, para considerar tanto a situação em que $f(x_i) - y_i > \varepsilon$ como $y_i - f(x_i) > \varepsilon$ (Schlkopf et

al., 2001). A Equação 60 mostra as restrições que devem ser observadas para alcançar a função objetivo (Vapnik, 1999), destacando-se que ξ e ξ^* podem ser nulos no caso do módulo do erro ser inferior a ε.

$$(a) \quad f(\vec{x}_i) - y_i \leq \varepsilon + \xi_i$$
$$(b) \quad y_i - f(\vec{x}_i) \leq \varepsilon + \xi_i^*$$
$$(c) \quad \xi_i, \xi_i^* \geq 0$$

Equação 60 – Restrições da Função Objetivo

Como os problemas do mundo real nem sempre podem ser representados diretamente por um modelo linear como o da Equação 57, é possível utilizar o mesmo artifício empregado na classificação SVM: mapear o espaço de entrada original para um espaço de características diferente, com dimensionalidade mais alta, onde o modelo se torne linear. Entretanto, assim como na classificação, o mapeamento não precisa ser efetuado explicitamente, se um kernel k for utilizado (Schlkopf et al., 2001). Assim, aplicando o kernel, a função objetivo é alcançada a utilização de um dos multiplicadores de Lagrange, definindo assim a função linear mostrada na Equação 61.

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) \cdot k(x_i, x) + b$$

Equação 61 – Função de Regressão

A demonstração de como a Equação 61 é obtida a partir da função objetivo é omitida neste trabalho. Para considerações mais aprofundadas sobre este tema, é oportuno considerar que se trata de uma generalização do SVM de margem suave, como detalhado em (Vapnik, 1999; Schlkopf et al., 2001).

Como visto, a classificação SVM depende somente de um subconjunto de dados (os vetores de suporte), ignorando as amostras que se encontram fora da margem de classificação. Analogamente, a regressão SVM depende de um subconjunto dos dados, porque a função de perda ignora quaisquer padrões que estejam suficientemente próximos (dentro de uma distância ε) à função estimada (Figura 26).

A Figura 27 mostra um diagrama esquemático do SVR, considerando o kernel $k(x_i, x)$ como o produto interno de x_i e x. O mapeamento denotado por $\Phi(x)$, é mostrado por razões conceituais, uma vez que de acordo com a Equação 27, $k(x_i, x) = \langle \vec{x}_i, \vec{x} \rangle = \langle \Phi(x_i), \Phi(x) \rangle$ e, portanto, o mapeamento pode ser omitido.

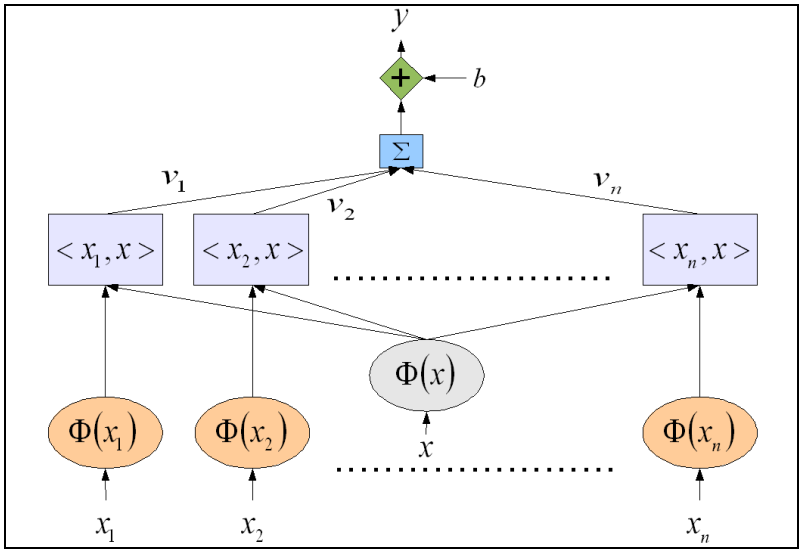

Figura 27 – Arquitetura da Regressão SVM

Como mostrado na Figura 27, a resposta para uma entrada desconhecida x é obtida em três etapas. Primeiro, são calculados os produtos internos de cada vetor de suporte com esta entrada. Na seqüência, estes produtos são combinados linearmente com os fatores $v_i = (\alpha_i^* - \alpha_i)$ e, finalmente, este resultado é somado ao termo constante b, tal como mostrado na Equação 61. Este processo é basicamente o mesmo que acontece em uma rede neuronal, excetuando-se pelo fato que no SVM os pesos sinápticos das unidades de entrada são constituídos por um subconjunto dos dados de treinamento (os vetores de suporte). Na Figura 27, isso denotado utilizando-se o índice n para o total de padrões existentes, ao invés de m, o total de amostras. Na Equação 61, isso é denotado implicitamente, ao admitir que $\alpha_i^* - \alpha_i$ possa ser nulo para alguns valores de i.

É importante destacar que o SVM escolhe a função mais linear possível entre aquelas que predizem os dados originais com uma dada precisão. Conceitualmente, a linearidade é obtida somente no espaço de características; entretanto, é importante notar

que a função de regressão equivalente no espaço de entrada original tende a ser suave. Isso se deve ao fato de que os kernels podem impor essa suavidade através de operadores de regularização (Schlkopf et al., 2001).

5 ESTADO DA ARTE

Embora a predição de carga elétrica venha sendo explorada há muitos anos, o foco deste trabalho é uma nova área de pesquisa. Esta pesquisa, ao mesmo tempo em que considera o já consolidado campo de predição como cenário de fundo, propõe uma abordagem nova baseada na detecção de similaridades entre perfis de consumo para aprimorar o desempenho de algoritmos de predição conhecidos, tal como o PCarga (Oliveira, 2004).

De acordo com o período de previsão, abordagens diferentes são empregadas para prever a carga elétrica. Para a predição de curto prazo, os algoritmos neurais híbridos são escolhas quase universais, porque asseguram alta precisão e são capazes de manipular grandes quantidades de informação (Guo et al., 2004; Hong et al., 2005). Nos últimos anos, alguns trabalhos bem sucedidos têm proposto modelos híbridos empregando máquinas de vetores de suporte (SVM). Os resultados obtidos com tais modelos são equivalentes ou superiores a aqueles obtidos com o uso de redes neurais artificiais (RNAs) (Tao et al., 2004; Hong et al., 2005; Niu et al., 2005; Guo et al., 2006).

Assim como no caso dos modelos básicos, o desempenho dos modelos híbridos é fortemente influenciado pela natureza dos dados de entrada. Uma vez que, em geral, dados brutos não são muito expressivos, uma parcela significativa de recursos computacionais é gasta para transformá-los adequadamente. No caso dos algoritmos híbridos, módulos específicos processam os dados de entrada antes que uma RNA ou SVM possa realizar inferências. Omitir esse passo, conhecido como *pré-processamento*, implica em degradar o desempenho do modelo de predição (Tao et al., 2004; Guo et al., 2004; Oliveira, 2004; Hong et al., 2005). Na realidade, o desempenho geral do modelo depende mais do pré-processamento do que da arquitetura da RNA (ou SVM).

A Figura 28 mostra um diagrama que mostra as etapas normalmente existentes nos sistemas para gerar preditores de carga, sendo que as etapas relacionadas ao pré-processamento estão destacadas por uma linha tracejada. Os trabalhos da área propõem abordagens ligeiramente diferentes entre si mas, em geral, implementam o esquema

mostrado nesta figura. Como pode ser percebido, o módulo de pré-processamento é implementado por quatro etapas distintas: (a) a seleção de parâmetros, (b) um algoritmo de avaliação, (c) um modelo de avaliação e (d) um critério de parada.

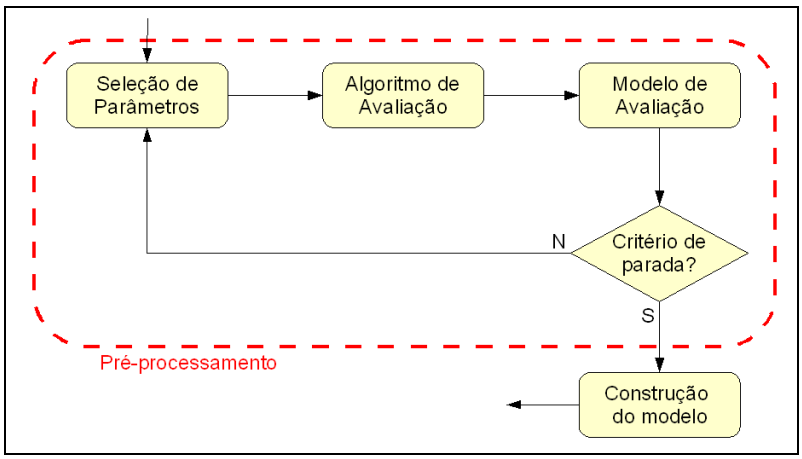

Figura 28 – Etapas na Construção de um Modelo Preditor

No contexto da Figura 28, a seleção de parâmetros é responsável por gerar iterativamente conjuntos diferentes com as variáveis de entrada, os quais possuem sua relevância preditiva testada depois de gerados. Se for constatado que um determinado conjunto é relevante, então ele é considerado um conjunto de preditores (ou parâmetros) válido.

A cardinalidade do conjunto preditor é uma das incógnitas a serem descobertas na seleção de parâmetros. Como na maior parte das situações práticas as variáveis de entrada são consideravelmente redundantes (Tao et al., 2004), é mesmo possível determinar dois conjuntos preditores de cardinalidade distinta. No entanto, alguns trabalhos ignoram esta peculiaridade da predição de carga e utilizam um conjunto fixo de variáveis (Iyer et al., 2003; Niu et al., 2005), o que pode reduzir a acurácia em algumas aplicações de predição de carga (Tao et al., 2004; Oliveira, 2004; Hong et al., 2005).

O algoritmo de avaliação define como os modelos de avaliação são gerados. Como em geral há muitas variáveis de entrada disponíveis, a seleção de parâmetros é intrinsecamente lenta. Portanto, é importante que o algoritmo de avaliação assegure uma rápida convergência, para diminuir o *overhead* geral na geração do preditor.

O modelo de avaliação é criado pelo algoritmo de avaliação utilizando o conjunto de parâmetros gerados no início do fluxo mostrado na Figura 28. O objetivo do modelo de avaliação é testar os conjuntos de parâmetros em condições de operação realísticas, de tal forma que o desempenho obtido com ele seja o desempenho esperado do modelo definitivo. A idéia, mostrada na Figura 29, é utilizar um determinado conjunto de preditores para alimentar o modelo de avaliação. Se este conjunto permitir que o modelo de avaliação prediga o passado recente a partir do passado remoto com razoável desempenho, então o modelo definitivo será capaz de prever o futuro com base no presente com desempenho semelhante.

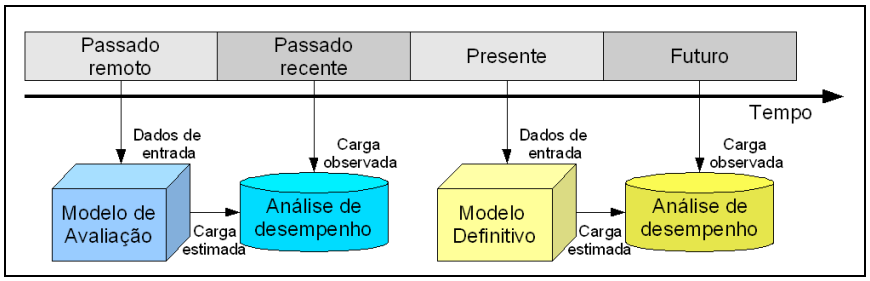

Figura 29 – Analogia entre os Modelos de Avaliação e Definitivo

Conforme mostrado na Figura 28, diversos modelos de avaliação podem ser gerados iterativamente até que um conjunto de preditores válidos seja encontrado (Tao et al., 2004; Oliveira, 2004; Hong et al., 2005). Como o número de iterações pode ser muito grande, é possível a utilização de modelos de avaliação mais simples que os modelos definitivos (Tao et al., 2004). Com isso, as iterações se tornam mais breves e os conjuntos de parâmetros gerados podem ser testados mais rapidamente.

O critério de parada normalmente é uma função de desempenho: se o conjunto de parâmetros testado permite o modelo de avaliação alcançar um patamar de desempenho previamente definido como satisfatório, então o pré-processamento está concluído; caso contrário, um novo conjunto de parâmetros deve ser gerado e testado. No caso dos modelos neurais, que empregam o princípio da minimização do risco empírico, é preciso tomar algumas precauções contra o super ajuste, tal como a validação cruzada. A não observância dessas precauções pode conduzir à perda de generalização dos modelos (Haykin, 1998; Duda et al., 2000). No caso das máquinas SV, a possibilidade de supera-

juste é minimizada através do controle da complexidade do modelo (minimização de risco estrutural) (Vapnik, 1998; Vapnik, 1999, Schlkopf et al., 2001).

Muitos trabalhos empregam soluções diferentes para o pré-processamento em modelos híbridos. Oliveira (2004) propôs a utilização de um algoritmo genético (AG) para realizar a seleção de parâmetros. O AG tenta formar um conjunto de preditores através de uma busca heurística inspirada na biologia evolucionária, utilizando técnicas como herança, mutação, seleção e recombinação (crossover). O critério de parada é fornecido pela função de *fitness* do AG, associada ao desempenho do modelo de avaliação. Como somente as variáveis com relevância preditiva são fornecidas para o modelo definitivo, este esquema assegura uma precisão maior e um tempo de convergência menor. Por outro lado, o AG exige um cluster com vários computadores para rodar o pré-processamento eficientemente.

Outras abordagens análogas são propostas por Hong et al. (2005) e Tao et al. (2004). Em Hong et al. (2005), a seleção de parâmetros é feita com um algoritmo de têmpora simulada (*Simulated Annealing Algorithm*), que é uma generalização do método de Monte Carlo para minimizar funções multivariadas. Já em Tao et al. (2004), a seleção de parâmetros é efetuada mediante uma estratégia de busca em árvore denominada método de busca flutuante (*Floating Search Method*). Nestas duas abordagens, os modelos definitivos são construídos com máquinas SV.

Guo et al. (2004) foge do esquema geral mostrado na Figura 28 ao sugerir um método de seleção de parâmetros baseado no uso de PCA (*Principal Components Analysis*, Análise de Componentes Principais). A PCA explora as correlações estatísticas existentes nos dados de entrada, produzindo um conjunto de variáveis transformadas (os *componentes principais*) que são combinações lineares das variáveis originais (Johnson et al., 2002; Rencher, 2002).

Os fatores principais são ordenados por ordem de importância, a qual é percebida como o impacto na variabilidade total dos dados de entrada. Se nestes dados forem percebidas correlações muito fortes, é possível utilizar um pequeno número de fatores para reproduzir a variabilidade original. Assim, um conjunto com n variáveis de entrada poderia ser substituído por m componentes principais, onde $m \leq n$. Por exemplo, nos

estudos conduzidos durante esta pesquisa, verificou-se que na subestação ILHA-CENTRO durante o inverno (perfil de consumo ICO_INV), percebeu-se cinco componentes principais conseguiam reproduzir cerca de 85% da variabilidade total observada num conjunto de 104 variáveis. Porém, como as grandezas meteorológicas e elétricas são dinâmicas, as correlações entre elas também são dinâmicas; portanto, a PCA pode conduzir a inconsistências se o período de predição for muito longo. No entanto, os resultados mostrados em (Guo et al., 2004) indicam que o método é especialmente poderoso quando aplicado a predições de até 24 horas adiante.

Iyer et al. (2003) desenvolveu um sistema neural difuso para uma aplicação que depende da predição de carga: a predição do preço da energia no mercado *spot*. O mercado spot trata a energia como uma commodity, negociada em transações à vista e entregue imediatamente. Em geral, as concessionárias distribuidoras compram energia das geradoras e transmissoras em leilões nos mercados atacadistas de energia (as chamadas câmaras de comercialização) com grande antecedência. Essa energia é entregue aos consumidores finais ao longo de um grande período de tempo. Se as previsões de demanda de consumo no longo prazo estiverem incorretas, as distribuidoras podem ficar com energia sobrando ou faltando no curto prazo, o que as obriga a negociar esta diferença no mercado spot quase em tempo real. Esta característica dinâmica é acentuada pela atuação de especuladores, que atuam no mercado spot da mesma forma como numa bolsa de valores. Como resultado, os preços de compra e venda de energia oscilam muito, dependendo do consumo de energia no curto prazo.

Para predizer o valor da energia no mercado spot, o sistema proposto por Iyer et al. (2003) emprega uma RNA alimentada por dados oriundos de duas fontes distintas, como mostrado na Figura 30: uma série histórica de preços e um módulo preditor difuso que prevê a demanda de carga no curto prazo. O preditor difuso, por sua vez, é alimentado por um conjunto fixo de variáveis (*dia*, *hora*, *estação* e *temperatura*) e prediz a carga elétrica como um valor discreto de consumo; ou seja, um tipo de carga, que pode ser *alto*, *médio* ou *baixo*. Os resultados mostrados neste trabalho são promissores, mas como as regras de inferência do módulo difuso são implementadas manualmente, uma aplicação em escala industrial poderia ser tecnicamente inviável.

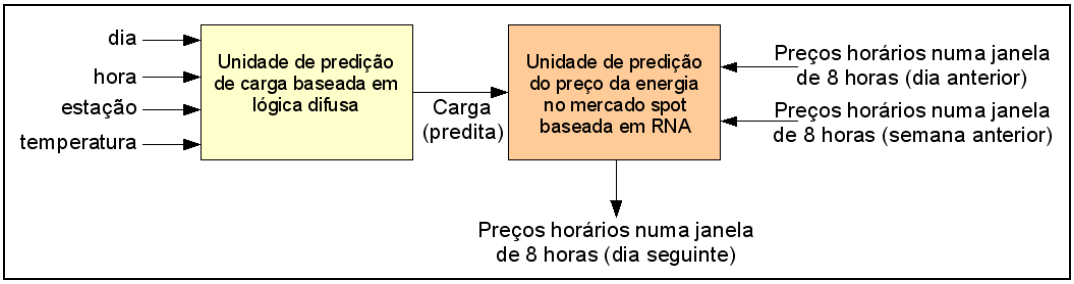

Figura 30 – Sistema Neural Difuso para a Predição do Preço da Energia no Mercado Spot

Niu et al. (2005) adotou um sistema que guarda algumas semelhanças com a proposta de Iyer et al. (2003), mas no contexto de predição de carga de curto prazo. Nesta solução, mostrada esquematicamente na Figura 31, o consumo de energia é classificado em três níveis (tipos) discretos: *alto*, *médio* e *baixo*. Uma RNA analisa as preditoras (valores históricos da carga, *temperatura*, *umidade*, *dia da semana* e *feriado*) e prediz o nível de consumo. Para cada tipo de consumo, existe uma máquina SV específica, treinada para predizer carga desse tipo.

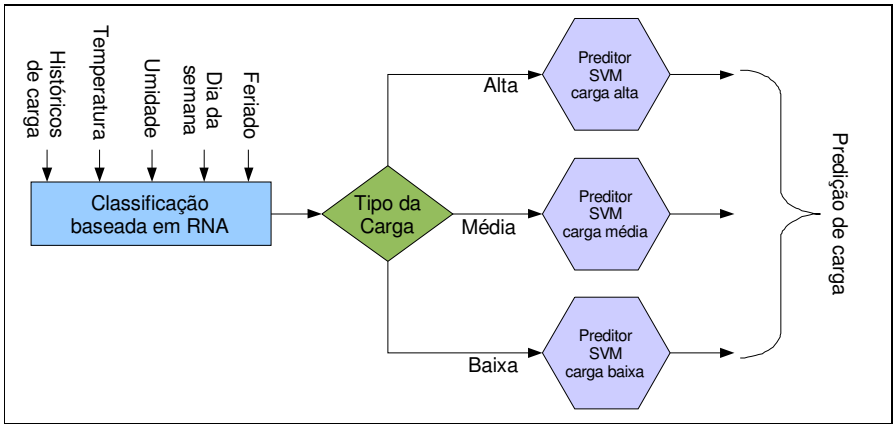

Figura 31 – Modelo Preditor Híbrido RNA-SVM

Na descrição dos resultados, Niu et al. (2005) assevera que a utilização da classificação neuronal permite a composição mais racional dos conjuntos de treinamento; isto é, possibilita a criação de conjuntos *balanceados*. O fato é que a utilização de três modelos preditores limita a variabilidade da carga que cada um deles deve predizer e, conseqüentemente, os resultados tendem a ser melhores do que seria possível obter com

modelos neuronais ou SV básicos. Esta solução é assaz criativa, mas é importante destacar que, da forma como foi concebida, a classificação da carga é feita com um conjunto pré-definido de variáveis, as quais também são utilizadas para alimentar os preditores. Assim, é possível que algumas variáveis relevantes sejam omitidas, o que poderia restringir o desempenho da predição.

Embora a arquitetura de predição de carga mostrada no presente trabalho seja original, ela utiliza uma forma generalizada da classificação de carga. No entanto, diferentemente da proposta de Niu et al. (2005), onde a carga dos perfis é categorizada em níveis discretos, no presente trabalho os próprios perfis são categorizados antes que seu preditor seja construído. Portanto, trata-se de abordagens muito distintas.

Como esta seção mostra, o pré-processamento é uma característica comum nas soluções de predição, exceto quando são utilizados modelos de séries temporais tal como o proposto por Guo et al. (2006). O pré-processamento assegura um desempenho superior à predição tanto em termos de precisão como convergência. Não obstante, o conceito de similaridade entre perfis pode ser aprimorar ainda mais a predição através da reciclagem do conhecimento existente nos preditores já consolidados. Na prática, um critério de similaridade poderia facilitar a inicialização dos parâmetros livres das RNAs e subsidiar a escolha dos preditores ótimos de um novo perfil de consumo.

A inicialização dos parâmetros em uma RNA é um gigantesco campo de estudo por si só, mas poderia ser facilitado dentro do contexto deste trabalho pela aplicação de uma heurística muito simples: *preditores de perfis similares tendem a serem parecidos*. Como destacado no Capítulo 1, um único preditor seria adequado para dois perfis que sejam idênticos. Ainda que perfis *idênticos* sejam conceitos ideais, perfis *similares* são comuns e compartilham algumas características. No caso das RNAs, a similaridade entre perfis leva a superfícies de erros e pontos críticos semelhantes. Assim, é mais fácil treinar uma RNA analisando as já consolidadas do que fazê-lo sem nenhum conhecimento a priori.

Os benefícios da aplicação do critério de similaridade são demonstrados através dos resultados empíricos mostrados no Capítulo 6. Além disso, são desenvolvidas duas ferramentas gráficas exploratórias – o espaço de características e o espaço causal, tam-

bém mostradas no Capítulo 6 – foram desenvolvidas para suportar tal critério, além de permitir inferências mais profundas sobre perfis de consumo e o seu comportamento.

6 Método de Otimização da Predição de Carga

O método de otimização proposto na seção 1.3 baseia-se na comparação entre perfis de consumo, os quais determinam o comportamento da carga elétrica numa região de consumo face à ação de alguns fatores de influência (as preditoras). É importante destacar que uma região de consumo é um critério de delimitação geográfica, enquanto que um perfil é um processo estocástico que descreve a carga elétrica numa região em um determinado instante.

O ganho obtido com o método consiste no reaproveitamento do conhecimento incorporado nos preditores de carga construídos e homologados. Os capítulos 1, 3 e 5 mostram que tal conhecimento consiste na modelagem das relações causais entre a carga elétrica e as preditoras. Especificamente, o Capítulo 5 descreve a seleção de parâmetros como um processo iterativo que exige um alto custo computacional (Oliveira, 2004), dado que a adequação do modelo deve ser conferida a cada iteração (Haykin, 1998; Vapnik, 1999; Duda et al., 2000).

Para que o conhecimento dos preditores consolidados seja reciclado, ele deve ser armazenado antes que novos modelos sejam construídos. A Figura 32 mostra a primeira etapa do método, que é a construção de uma base de conhecimento com os dados pertinentes a cada preditor (e, conseqüentemente, a cada perfil).

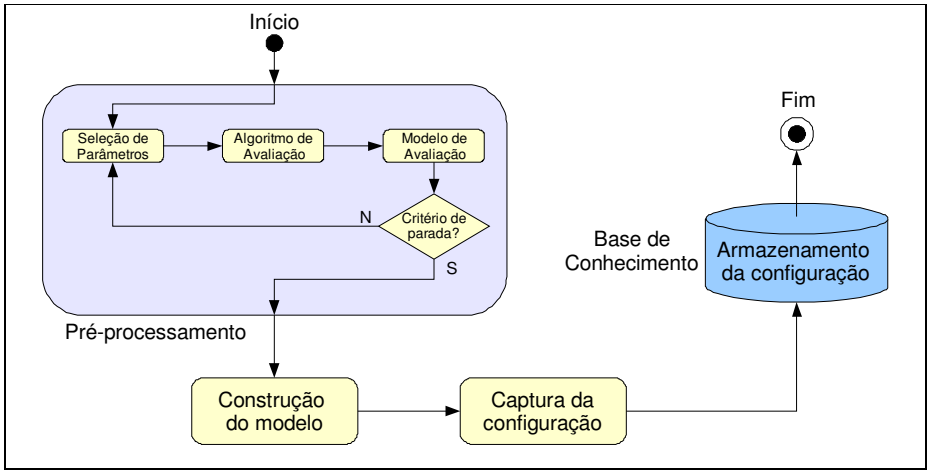

Figura 32 – Extraindo o Conhecimento dos Preditores de Carga

Inicialmente, os modelos são construídos a partir de uma estratégia de pré-processamento existente, como as discutidas no Capítulo 5. Após a construção do modelo preditor, sua configuração é capturada e armazenada na Base de Conhecimento.

A Base de Conhecimento armazena as preditoras, os modelos preditores e os vetores de características dos perfis. A utilização deste conhecimento exige um critério de similaridade para determinar qual configuração armazenada pode acelerar a criação do preditor de um perfil novo.

A Figura 33 mostra um diagrama onde a estratégia de otimização é mostrada, utilizando a arquitetura mostrada no Capítulo 1 (seção Metodologia). Como pode ser notado, a otimização proposta não substitui o pré-processamento (Figura 28), mas o complementa, fornecendo informações a priori que podem acelerar a seleção de parâmetros e a convergência do modelo. Além disso, a Base de Conhecimento continua armazenando as configurações dos modelos novos, sendo enriquecida toda vez que um novo preditor é criado.

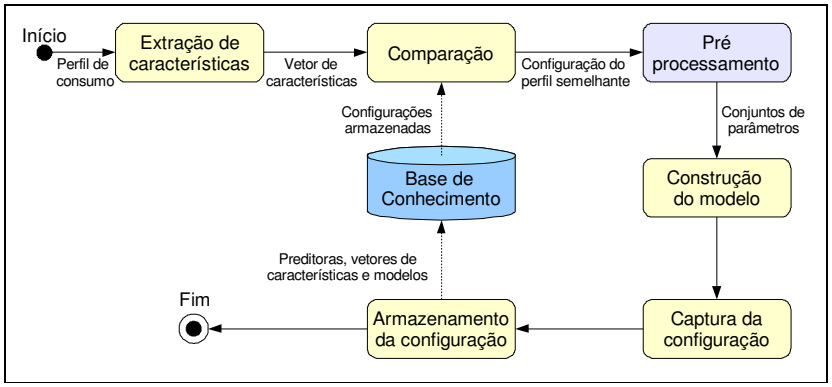

Figura 33 – Método de Otimização

Considere-se, por exemplo, o pré-processamento do PCarga (Oliveira, 2004), que utiliza um algoritmo genético (AG) para determinar o conjunto ótimo de preditores. A arquitetura geral do PCarga enquadra-se na Figura 32, e prevê um processo iterativo (o AG) onde diversos conjuntos de parâmetros são testados quanto à sua relevância preditiva. Um conjunto de parâmetros inicial, denominado *população inicial*, vai sendo modificado iterativamente através de operações evolutivas (*herança*, *mutação*, *seleção* e *recombinação*) até que o desempenho do modelo de avaliação seja considerado adequado. Com o método descrito neste trabalho, a população inicial seria dada pelos parâme-

tros do preditor do perfil mais semelhante. Assim, dependendo do grau de similaridade existente, poderiam ser necessárias menos iterações para determinar os parâmetros de um preditor novo.

Como a Figura 33 mostra, a primeira etapa da otimização é a extração de características dos perfis de consumo novos. A extração gera um vetor de características que é uma forma de representação mais conveniente do que os conjuntos amostrados que compõem os perfis de consumo. De posse deste vetor, é possível fazer uma busca eficiente na Base de Conhecimento, localizando o perfil armazenado que mais se assemelha ao perfil novo, bem como o modelo associado a este último. A partir daí, é possível construir um preditor novo a partir de um outro já consolidado e de um conjunto inicial de preditores. Com esta estratégia, espera-se que a construção de preditores novos se torne, em média, uma tarefa computacional menos onerosa.

Uma outra vantagem dos vetores de características é a facilidade de representação geométrica num espaço de características. As séries históricas de variáveis meteorológicas e elétricas que descrevem os perfis de consumo são usualmente representadas como histogramas ou gráficos de dispersão (Montgomery et al., 2001; Berry et al., 2004). Embora esta representação seja útil para determinados tipos de análise, ela não favorece a comparação entre os perfis de consumo. Com a definição de um espaço de características, muitas ferramentas úteis podem ser empregadas com esta finalidade, tal como a análise de agrupamentos. Dentro deste cenário, um espaço de características deve possuir as seguintes propriedades:

a) Perfis de consumo similares devem estar topologicamente mais próximos entre si do que de perfis dissimilares, formando agrupamentos.

b) A formação dos agrupamentos possui uma interpretação semântica consistente.

A propriedade (a) é pré-requisito para a análise de agrupamentos, a qual depende sobremaneira do arranjo topológico das amostras no espaço de características. A forma como a distância é calculada (distância euclidiana, euclidiana quadrática, Manhattam, Chebychev, entre outras), bem como as regras de amalgamação (ligação simples, completa, média par-grupo ponderada/não-ponderada, entre outras), podem variar de acordo

com o ambiente e os objetivos pretendidos (Duda et al., 2000). Todavia, qualquer que seja a configuração, a similaridade entre as amostras deve ser sempre uma função inversa da distância medida entre elas. Como conseqüência direta deste corolário, espera-se que as distâncias observadas entre as amostras de um mesmo agrupamento sejam significativamente menores do que as observadas entre amostras de agrupamentos diferentes (Hand et al., 2001).

A propriedade (b) é essencial à análise exploratória, tanto para testar hipóteses elaboradas a priori como para gerar hipóteses orientadas a dados (*data-driven hypothesis generation*) (Hand et al., 2001). Uma interpretação semântica consistente dos agrupamentos formados no espaço de características só é possível quando a extração de características captura aspectos relevantes dos perfis de consumo. Caso os critérios utilizados para a extração de características sejam precários, os agrupamentos não serão significativos, o que tornaria a comparação entre os perfis inconsistente.

6.1 Extração de Características

Em geral, os modelos de predição de carga podem ser divididos em duas categorias (Makridakis et al., 1997): *séries temporais*, baseada somente nos valores atuais e históricos da carga, e *explanatórios*, que utiliza extensivamente variáveis extrínsecas ao sistema elétrico, tais como radiação solar, temperatura e outros elementos climáticos (Guo et al., 2004; Oliveira, 2004). Em ambos os casos, a cardinalidade do conjunto de variáveis é relativamente grande, facilmente excedendo uma centena. Desta forma, uma opção natural para a representação das subestações seria a aplicação de técnicas multivariadas como a análise de componentes principais (PCA), visando à redução da dimensionalidade. Com efeito, o emprego da PCA pode ser vantajoso na predição de curto prazo (onde o período de predição varia de poucos minutos a algumas horas adiante), tal como demonstrado por (Guo et al., 2004). Entretanto, dentro do escopo desse trabalho, algumas premissas necessárias para a obtenção de resultados consistentes nem sempre são observadas nas regiões de consumo.

A carga elétrica de uma subestação típica não adere aos conceitos de estacionariedade e normalidade, que são pré-requisitos à PCA e muitas outras técnicas multivariadas. A Figura 34 mostra como a variância da carga se comporta durante o inverno em um perfil de consumo urbano (centro de Florianópolis, subestação Ilha-Centro).

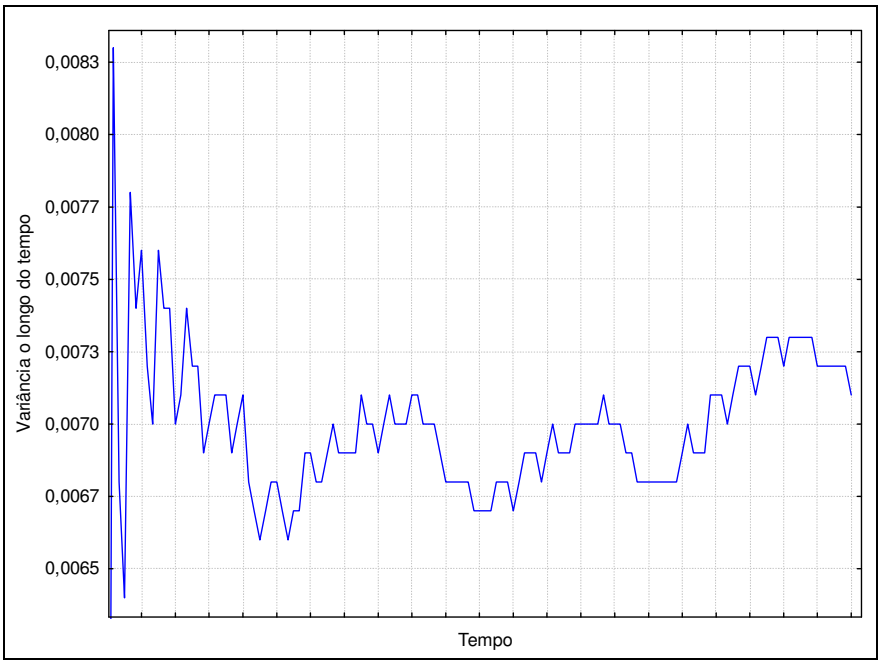

Figura 34 – Não Estacionariedade da Carga

Cada ponto da curva mostrada na Figura 34 é calculado incrementalmente; ou seja, o primeiro valor é a variância das primeiras *m* amostras, o segundo é a variância das primeiras $2 \times m$ amostras e assim por diante. Para um sistema estacionário, a curva se estabilizaria, tendendo a um valor constante. Entretanto, o comportamento mostrado na figura é outro – a variância oscila consideravelmente ao longo do tempo sem convergir para um valor definido, o que é típico de sistemas não-estacionários.

Por outro lado, a normalidade do mesmo sinal é analisada com o uso do gráfico de probabilidade normal mostrado na Figura 35. *Uma vez que a premissa de normalidade conduz a uma série inferências que podem ser espúrias caso a distribuição subjacente não se distribua normalmente* (Montgomery et al., 2001), a análise da Figura 35 é altamente significativa. Pequenos desvios da premissa de normalidade não afetam a aplicação das técnicas multivariadas; entretanto, se esses desvios forem consideráveis, os resultados serão inconsistentes.

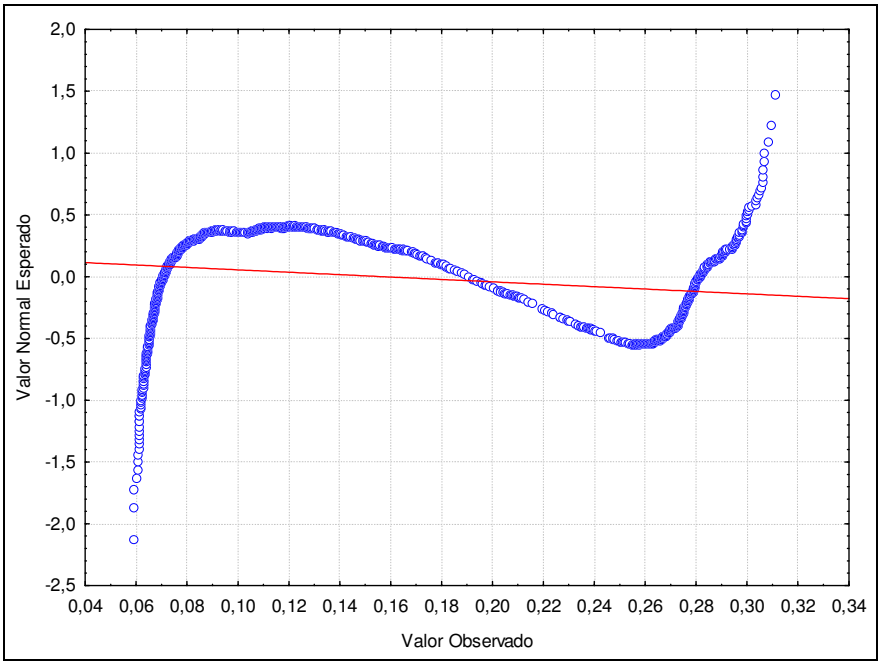

Figura 35 – Gráfico de Probabilidade Normal da Carga Elétrica

Um gráfico de probabilidade normal é um diagrama de dispersão aonde os resíduos do sinal testado (dispersão de cada amostra em relação à média) são plotados contra os resíduos esperados caso a distribuição fosse realmente normal. Portanto, se a premissa de normalidade for verdadeira, a curva formada se aproximará da identidade. No gráfico da Figura 35, os resíduos observados da carga elétrica ativa (em azul) são ordenados e traçados contra os valores esperados dos resíduos considerando uma distribuição normal. Para efeitos de comparação, o mesmo gráfico mostra um gabarito para o padrão de normalidade, indicado pela reta traçada em vermelho. Para salientar o desvio da normalidade, a tendência linear deste gráfico é removida, formando o que a literatura de nomina de *probabilidade normal sem tendências* (*detrended normal probability*).

Conforme pode ser constatado pela análise da Figura 35, os resíduos formam uma curva acentuada, diferindo consideravelmente do padrão esperado para a normalidade. Dessa forma, não é possível assumir que a distribuição dessa grandeza é gaussiana (Montgomery et al., 2001). Essa conclusão é corroborada pela análise da Figura 36, que mostra um histograma discretizado da carga elétrica. Claramente, há um desvio do padrão esperado para a normalidade.

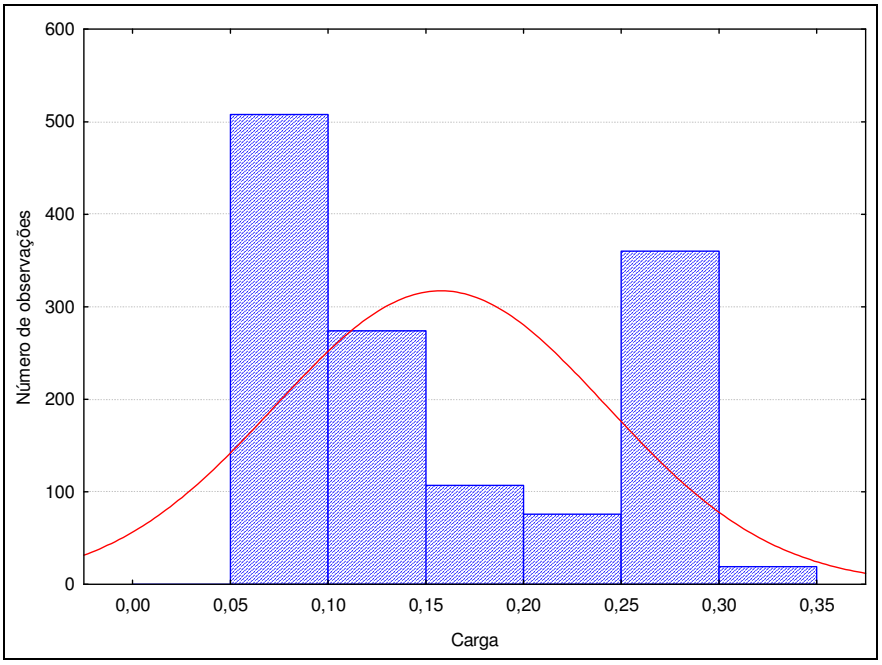

Figura 36 – Histograma da Distribuição da Carga

É importante destacar que as não-idealidades mostradas nos gráficos das figuras Figura 34, Figura 35 e Figura 36 também se aplicam às variáveis preditoras. Ora, a carga elétrica é o resultado da contribuição de muitas variáveis idêntica e independentemente distribuídas, tal como demonstrado empiricamente em Tao et al. (2004), Guo et al. (2004), Oliveira (2004) e Hong et al. (2005). Por esta razão, em princípio seria esperada uma distribuição normal para a carga, tal como assegurado pelo Teorema do Limite Central (Johnson et al., 2002). De fato, existem muitos sistemas gaussianos dinâmicos que apresentam um comportamento variante no tempo. Então, a despeito do comportamento dinâmico mostrado na Figura 34, a carga elétrica deveria ser distribuída normalmente. Todavia, face o teste de normalidade da Figura 35, esta hipótese não pode ser aceita.

É possível que massas de dados maiores pudessem reproduzir um comportamento normal. No entanto, a coleta de dados num perfil é rigidamente limitada no tempo e no espaço: um perfil de consumo é transitório e está restrito a uma área geográfica bem definida. De todo modo, ainda que fosse possível empregar amostras maiores, o comportamento dinâmico destacado na Figura 34 só pode ser contornado mediante a utilização de amostras menores, onde a oscilação da variância não seja expressiva (Oppe-

nheim et al., 1999). Isto estabelece um paradoxo que restringe a aplicação das técnicas multivariadas normalmente empregadas em problemas correlatos.

Para suplantar estas limitações, esta pesquisa explorou uma propriedade dos modelos de regressão (estimadores), relacionada à sua capacidade de modelar um sistema cujas características são desconhecidas a priori (Haykin, 1998). Algoritmos como o SVM (Support Vector Machine) e o backpropagation extraem conhecimento de um sistema através de processos iterativos que minimizam uma função de erro (Duda et al., 2000; Schlkopf et al., 2001). Essas funções de erro envolvem a relação entre as variáveis de entrada e a saída desejada, que é específica para cada sistema; portanto, os estimadores tornam-se ad hoc, tal como uma espécie de assinatura do sistema.

No presente trabalho, as entradas são as preditoras, a saída é a carga elétrica ativa e o sistema é o perfil de consumo. Do ponto de vista conceitual, como discutido no Capítulo 3, estimadores e preditores são semelhantes. Entretanto, no contexto deste trabalho, os preditores são mais custosos em termos de processamento do que os estimadores, principalmente devido à seleção de variáveis com relevância preditiva (Tao et al., 2004; Oliveira, 2004; Hong et al., 2005).

Seja um determinado estimador ξ_x especificamente construído para um perfil de consumo desconhecido ρ_x. Como as propriedades de ρ_x não são conhecidas, ξ_x emprega todas as variáveis disponíveis como entradas, razão pela qual é denominado de *estimador leigo*. Como discutido no Capítulo 1, esta abordagem acrescenta um erro à resposta do estimador porque somente um subconjunto de todas as variáveis disponíveis possui relevância preditiva. Entretanto, o poder de ξ_x para estimar carga não é o foco: o seu real objetivo é detectar similaridades entre os perfis de consumo, de forma que o pré-processamento possa ser acelerado (Figura 33), otimizando a predição de carga. Por outro lado, como é justamente a seleção de parâmetros que torna onerosa a construção de um preditor de carga (Tao et al., 2004; Oliveira, 2004; Hong et al., 2005), os estimadores leigos são computacionalmente fáceis de serem construídos. Desta forma, a sua existência torna-se amplamente justificada porque eles podem acelerar a construção de preditores de carga.

Se ξ_x é empregado para estimar a carga de outro perfil ρ_y e o desempenho observado é satisfatório, assume-se que ρ_x e ρ_y são similares; caso contrário, eles são dissimilares. Este raciocínio é baseado no fato que um estimador, assim como um preditor, é um modelo *ad hoc* de um sistema – se outro sistema é modelado pelo mesmo estimador com um desempenho comparável, então é plausível assumir que ambos os sistemas têm algo em comum (ou seja, são *similares* de alguma forma).

Para que este critério de similaridade seja considerado válido dentro dos objetivos deste trabalho, é necessário comprovar que regiões de consumo similares possuem conjuntos similares de preditoras. Isto pode ser expresso através das seguintes hipóteses:

H_A – O desempenho de um conjunto de estimadores de carga elétrica, quando aplicado a um perfil de consumo, determina um vetor de características que o representa;

H_B – Dois perfis de consumo cujos vetores de características são similares compartilham conjuntos semelhantes de variáveis preditoras.

A Figura 37 mostra o desempenho de um conjunto de estimadores leigos ao processar um conjunto de perfis de consumo, produzindo curvas denominadas *curvas de desempenho*. Cada ponto dessas curvas é o desempenho de um estimador ao processar um perfil ou, mais precisamente, as variáveis de entrada de um perfil.

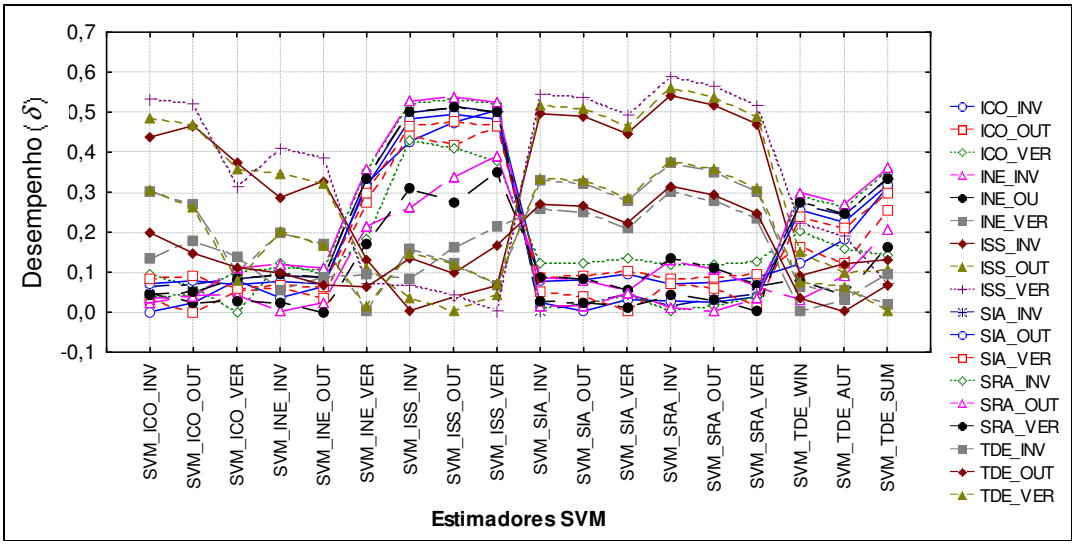

Figura 37 – Gráfico de Desempenho de Diversos Perfis de Consumo

Por se basear no desempenho dos estimadores leigos, este gráfico foi denominado de *gráfico de desempenho*. No eixo das abscissas, são mostrados os estimadores leigos, implementados com SVM, enquanto que o eixo das ordenadas mostra o erro médio quadrático (desempenho do estimador). Neste trabalho, o desempenho de um estimador é definido como o erro médio quadrático (*Root Mean Squared Error*, RMSE) entre os valores observados e estimados da carga. A Equação 62 mostra esta definição, onde δ é o RMSE, m é o número de observações, y_i é o valor observado da carga e $\hat{y}_i$ é o seu valor estimado.

$$\delta = \sqrt{1/m \cdot \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$$

Equação 62 – Desempenho de um Estimador Expresso pelo seu RMSE

Cada um dos estimadores leigos indicados na Figura 37 foi construído especificamente para um dos perfis de consumo processados. Assim, espera-se que o melhor desempenho esperado de cada estimador (ponto de mínimo da curva de desempenho) seja obtido quando ele processa o perfil para o qual foi construído.

As curvas de desempenho da Figura 37 se referem a 6 regiões de consumo do estado de Santa Catarina, a saber:

a) **Região Trindade (TDE)** – região litorânea de Florianópolis cuja carga é mista (residencial e comercial). Abastece também a Universidade Federal de Santa Catarina, que é o maior consumidor individual da ilha de Santa Catarina.

b) **Região Ilha-Norte (INE)** – região norte de Florianópolis, predominantemente constituída por residências e comércio de veraneio.

c) **Região Ilha-Centro (ICO)** – região central de Florianópolis, constituída por carga predominante comercial e um pequeno percentual residencial.

d) **Região Itajaí-Salseiros (ISS)** – região litorânea do Vale do Itajaí (cidade de Itajaí), com carga residencial predominantemente unifamiliar, comercial, industrial e portuária (porto de Itajaí).

e) **Região Sadia (SIA)** – região de consumo constituída pelo transformador da subestação Concórdia (cidade de Concórdia, no oeste do estado de Santa Catarina) que abastece a indústria de alimentos Sadia S/A.

f) **Região Seara (SRA)** – região de consumo do oeste de Santa Catarina (cidade de Seara), com carga residencial predominantemente unifamiliar, comercial e industrial.

Estas 6 regiões de consumo são analisadas durante três estações do ano de 2003: inverno (INV), outono (OUT) e verão (VER). Como geralmente cada estação do ano impõe um perfil de consumo diferente às regiões, elas foram divididas em 18 perfis de consumo, representadas pelas curvas de desempenho mostradas na Figura 37: ICO_INV, ICO_OUT, ICO_VER, INE_INV, INE_OUT, INE_VER, ISS_INV, ISS_OUT, ISS_VER, SIA_INV, SIA_OUT, SIA_VER, SRA_INV, SRA_OUT, SRA_VER, TDE_INV, TDE_OUT e TDE_VER.

A Figura 38 mostra um diagrama esquemático que relaciona alguns perfis aos seus estimadores. Como cada estimador modela um perfil de consumo, é possível comparar os perfis através dos respectivos estimadores. Assim, é o conjunto de estimadores que extrai as características de cada perfil, permitindo o uso de um recurso geométrico para representá-los convenientemente: *o espaço de características* dos perfis de consumo.

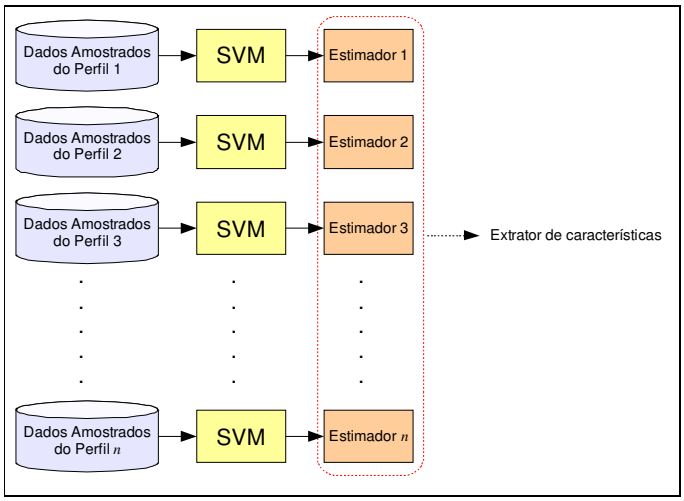

Figura 38 – Criação do Extrator de Características Utilizando SVM

Cada curva de desempenho da Figura 37 pode ser entendida como uma abstração de um perfil de consumo, útil na determinação de regularidades que podem otimizar a predição de carga. Como mostrado, o desempenho da estimação em cada perfil pode variar bastante, de acordo com o estimador utilizado. Dentro das hipóteses enunciadas, a similaridade entre as regiões de consumo é determinada pela semelhança entre as envoltórias de suas respectivas curvas de desempenho. Considere-se, por exemplo, o caso de dois perfis que mantêm grande similaridade entre si (SRA_OUT e SRA_VER) comparados com o caso de dois perfis dissimilares (ISS_OUT e SRA_INV).

No primeiro caso, mostrado na Figura 39, a similaridade é denotada por uma correlação forte entre as curvas de desempenho. Se traçados num espaço de características, espera-se que os perfis estejam muito próximos entre si. No segundo caso, mostrado na Figura 40, ocorre o inverso, sendo a dissimilaridade evidenciada pela anti-correlação das curvas. Portanto, num espaço de características, esses perfis deveriam estar afastados um do outro.

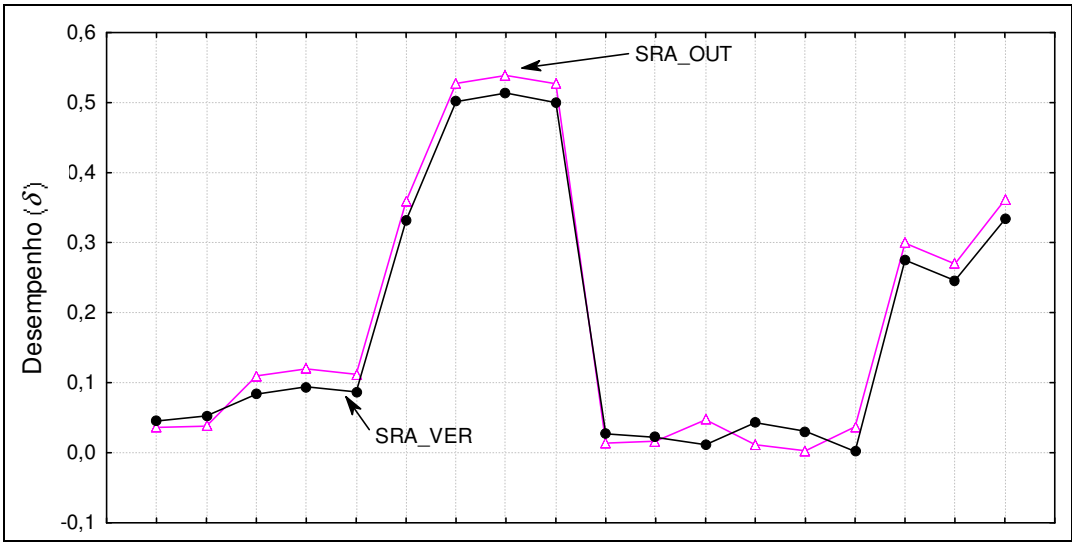

Figura 39 – Curvas de Desempenho de Dois Perfis de Consumo Semelhantes

As curvas mostradas nos gráficos de desempenho são bastante sugestivas. Embora a teoria que fundamenta a interpretação desses gráficos seja heurística e, portanto, de comprovação empírica, ela é corroborada por conceitos derivados da Teoria de Aprendizado Estatístico (Capítulo 3) e da estatística paramétrica, como discutido a seguir.

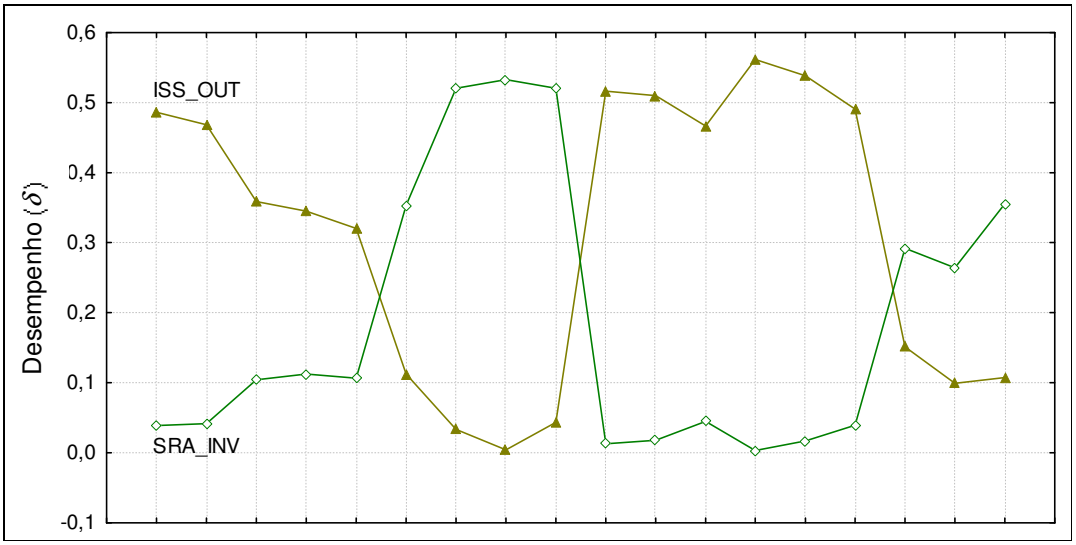

Figura 40 – Curvas de Desempenho de Dois Perfis de Consumo Não-Semelhantes

Analisando H_A

Por inspeção, parece haver três ou quatro padrões delineados na Figura 37, o que é corroborado pela análise da matriz de correlações dos vetores de características de cada região, mostrada na Tabela 2.

Tabela 2 – Correlações dos Vetores de Características de Cada Região de Consumo

	ICO INV	ICO OUT	ICO VER	INE INV	INE OUT	INE VER	ISS INV	ISS OUT	ISS VER	SIA INV	SIA OUT	SAI VER	SRA INV	SRA OUT	SRA VER	TDE INV	TDE OUT	TDE VER
ICO INV	1.00	0.98	0.88	0.95	0.92	-0.62	-0.82	-0.85	-0.89	0.95	0.95	0.95	0.94	0.94	0.95	-0.16	-0.33	-0.62
ICO OUT	0.98	1.00	0.93	0.94	0.95	-0.54	-0.80	-0.81	-0.85	0.94	0.94	0.94	0.93	0.93	0.94	-0.14	-0.26	-0.54
ICO VER	0.88	0.93	1.00	0.84	0.90	-0.35	-0.74	-0.73	-0.70	0.87	0.88	0.89	0.86	0.86	0.87	-0.12	-0.18	-0.35
INE INV	0.95	0.94	0.84	1.00	0.96	-0.42	-0.63	-0.68	-0.74	0.84	0.85	0.87	0.82	0.82	0.85	0.11	-0.09	-0.42
INE OUT	0.92	0.95	0.90	0.96	1.00	-0.32	-0.62	-0.64	-0.69	0.83	0.84	0.87	0.81	0.81	0.84	0.11	0.00	-0.33
INE VER	-0.62	-0.54	-0.35	-0.42	-0.32	1.00	0.86	0.88	0.90	-0.74	-0.72	-0.67	-0.76	-0.77	-0.73	0.74	0.88	1.00
ISS INV	-0.82	-0.80	-0.74	-0.63	-0.62	0.86	1.00	0.99	0.96	-0.94	-0.93	-0.90	-0.95	-0.96	-0.94	0.68	0.76	0.86
ISS OUT	-0.85	-0.81	-0.73	-0.68	-0.64	0.88	0.99	1.00	0.98	-0.95	-0.94	-0.91	**-0.96**	-0.96	-0.95	0.64	0.76	0.88
ISS VER	-0.89	-0.85	-0.70	-0.74	-0.69	0.90	0.96	0.98	1.00	-0.95	-0.94	-0.91	-0.96	-0.97	-0.95	0.56	0.71	0.90
SIA INV	0.95	0.94	0.87	0.84	0.83	-0.74	-0.94	-0.95	-0.95	1.00	1.00	0.99	1.00	1.00	1.00	-0.42	-0.55	-0.74
SIA OUT	0.95	0.94	0.88	0.85	0.84	-0.72	-0.93	-0.94	-0.94	1.00	1.00	0.99	0.99	0.99	1.00	-0.40	-0.53	-0.72
SIA VER	0.95	0.94	0.89	0.87	0.87	-0.67	-0.90	-0.91	-0.91	0.99	0.99	1.00	0.98	0.98	0.99	-0.34	-0.47	-0.67
SRA INV	0.94	0.93	0.86	0.82	0.81	-0.76	-0.95	**-0.96**	-0.96	1.00	0.99	0.98	1.00	1.00	1.00	-0.45	-0.58	-0.77
SRA OUT	0.94	0.93	0.86	0.82	0.81	-0.77	-0.96	-0.96	-0.97	1.00	0.99	0.98	1.00	1.00	**0.99**	-0.46	-0.58	-0.77
SRA VER	0.95	0.94	0.87	0.85	0.84	-0.73	-0.94	-0.95	-0.95	1.00	1.00	0.99	1.00	**0.99**	1.00	-0.42	-0.54	-0.74
TDE INV	-0.16	-0.14	-0.12	0.11	0.11	0.74	0.68	0.64	0.56	-0.42	-0.40	-0.34	-0.45	-0.46	-0.42	1.00	0.93	0.74
TDE OUT	-0.33	-0.26	-0.18	-0.09	0.00	0.88	0.76	0.76	0.71	-0.55	-0.53	-0.47	-0.58	-0.58	-0.54	0.93	1.00	0.88
TDE VER	-0.62	-0.54	-0.35	-0.42	-0.33	1.00	0.86	0.88	0.90	-0.74	-0.72	-0.67	-0.77	-0.77	-0.74	0.74	0.88	1.00

Na Tabela 2, as correlações das curvas mostradas na Figura 39 são destacadas em laranja, enquanto que as das curvas Figura 40, em verde. Basicamente, as correlações apresentam a mesma informação dos gráficos de desempenho na forma numérica.

Considere-se um conjunto de regiões de consumo ρ para o qual é construído um conjunto de estimadores leigos ξ, de acordo com a Equação 63 e a Equação 64. Em ambas as equações, *n* é o número total de estimadores e de regiões, sendo que ξ_x é o estimador de ρ_x para todo *x*=1, 2, ..., *n*.

$$\rho = [\rho_1, \rho_2, ..., \rho_n]$$

Equação 63 – Conjunto de Regiões de Consumo

$$\xi = [\xi_1, \xi_2, ..., \xi_n]$$

Equação 64 – Conjunto de Estimadores Leigos

A Equação 65 mostra $\delta_{x,y}$ como sendo o desempenho do estimador ξ_x quando aplicado à região ρ_y. Como cada estimador é construído especificamente para um perfil de consumo, espera-se que $\delta_{x,y}$ seja mínimo quando *x*=*y*, como observado na Figura 37.

$$\delta_{x,y} = \xi_x(\rho_y)$$

Equação 65 – O desempenho do Estimador ξ_x Aplicado à Região de Consumo ρ_y

Dado o problema de obter um modelo de predição π_w para ρ_w a partir do conjunto de estimadores ξ (definido na Equação 64), seja π_z (*z*=1, 2, ..., *n*) o modelo de predição de ρ_z, ξ_z o estimador de ρ_z, $\delta_{Z\max} = \delta_{z,z}$ o desempenho máximo de ξ_z e $\delta_{z,w}$ o desempenho de ξ_z aplicado a ρ_w. Seja também $\theta \in \Re$ uma constante arbitrariamente pequena, utilizada na Equação 66 e definida como limiar de similaridade.

$$\left|\delta_{z,w} - \delta_{Z\max}\right| \leq \theta$$

Equação 66 – Limiar de Similaridade entre ρ_w e ρ_z

Se a condição estabelecida na Equação 66 for atendida para um determinado *z*, então ρ_w e ρ_z são consideradas similares e a construção de π_w pode ser acelerada com as informações de π_z. Caso contrário, ρ_w e ρ_z são dissimilares e o critério é de pouca utilidade prática para este trabalho porque, embora se possa afirmar com segurança que ρ_w não se assemelha a ρ_z, não se pode dizer com o quê ρ_w se assemelha. Como conseqüência, não se dispõe de nenhuma informação extra para construir π_w.

Do exposto, fica estabelecida a necessidade de se utilizar mais figuras de mérito para a análise, como o vetor $\Delta_y \in \Re^n$ da Equação 67, constituído pelos desempenhos do conjunto de estimadores ξ aplicado a ρ_y. Dentro das premissas estabelecidas por H_A, Δ_y é o *vetor de características* de ρ_y e define sua curva de desempenho.

$$\Delta_y = \left(\delta_{1,y}, \delta_{2,y}, ..., \delta_{n,y}\right)$$

Equação 67 – O Vetor de Características da Região *y*

Tanto as curvas de desempenho como os vetores de características representam os perfis de consumo. De fato, a similaridade entre perfis pode ser mensurado tanto pela distância entre seus vetores de características como pela similaridade entre suas curvas de desempenho, tal como mostrado nas Figuras Figura 39 e Figura 40. Cada ponto de uma curva de desempenho determina uma dimensão do vetor de características sendo que, no caso deste trabalho em particular, as características são definidas num espaço $\Re^{18}$. A disponibilidade de mais estimadores permitiria incrementar a dimensionalidade deste espaço, ao mesmo tempo em que o torna menos esparso.

A Figura 41 mostra como as características de um perfil de consumo desconhecido são extraídos com o uso dos estimadores mostrados na Figura 38. Como pode ser visto, cada característica de um perfil é extraída por um estimador, formando o vetor descrito na Equação 67. Originalmente, um perfil é constituído por séries históricas (amostras da carga elétrica e das variáveis explanatórias) representadas numa matriz de alta cardinalidade e alta dimensionalidade; ou seja, grande número de linhas (amostras) e colunas (preditoras). A conveniência da abordagem desenvolvida neste trabalho consiste justamente em representar esta estrutura complexa através de um vetor de baixa

dimensionalidade n. Com o uso desta representação, a comparação entre perfis torna-se computacionalmente tratável, possibilitando a realização de inferências úteis.

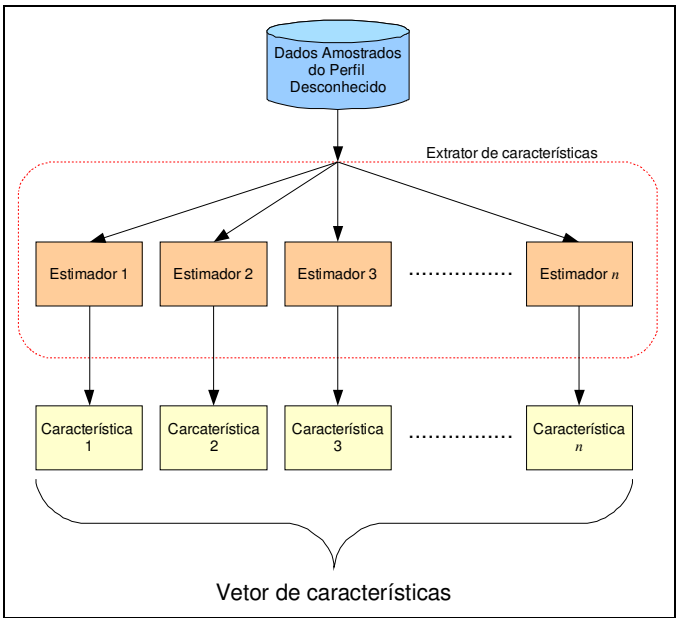

Figura 41 – Diagrama Esquemático do Extrator de Características

A Tabela 3 mostra o desempenho de todos os estimadores aplicados a todas as regiões de consumo, constituindo uma forma de representação alternativa a gráficos como os da Figura 37. A diagonal principal desta tabela armazena os menores valores de cada coluna, uma vez que nesta região os índices de linha e coluna são iguais e, como destacado anteriormente, espera-se que $\delta_{x,y}$ seja mínimo quando $x = y$.

Tabela 3 – Desempenho de Todos os Estimadores Aplicados a Todos os Perfis de Consumo

Δ_1	Δ_2	...	Δ_n
$\delta_{1,1}$	$\delta_{1,2}$	...	$\delta_{1,n}$
$\delta_{2,1}$	$\delta_{2,2}$	...	$\delta_{2,n}$
...	...	...	
$\delta_{n,1}$	$\delta_{n,2}$	...	$\delta_{n,n}$

O gráfico da Figura 42 mostra as curvas de desempenho da Figura 37 representadas como vetores no espaço de características. Nesse espaço, reduzido ao plano cartesiano com o uso de escalonamento multidimensional, nota-se formação de agrupamentos bem definidos e consistentes com a definição intuitiva de similaridade entre perfis. Por

exemplo, nas áreas onde há sazonalidade de consumo relacionada à temperatura, os perfis de consumo correspondentes às estações frias (outono e inverno) estão situados em agrupamentos diferentes da estação quente (verão).

Diante da análise efetuada e da evidência gráfica da Figura 42, verifica-se que a hipótese H_A é consistente, sendo aceita. Uma vez que o critério de similaridade proposto parece ser adequado para representar as regiões de consumo, resta demonstrar se os conjuntos de preditoras são semelhantes para regiões da mesma classe, como asseverado por H_B.

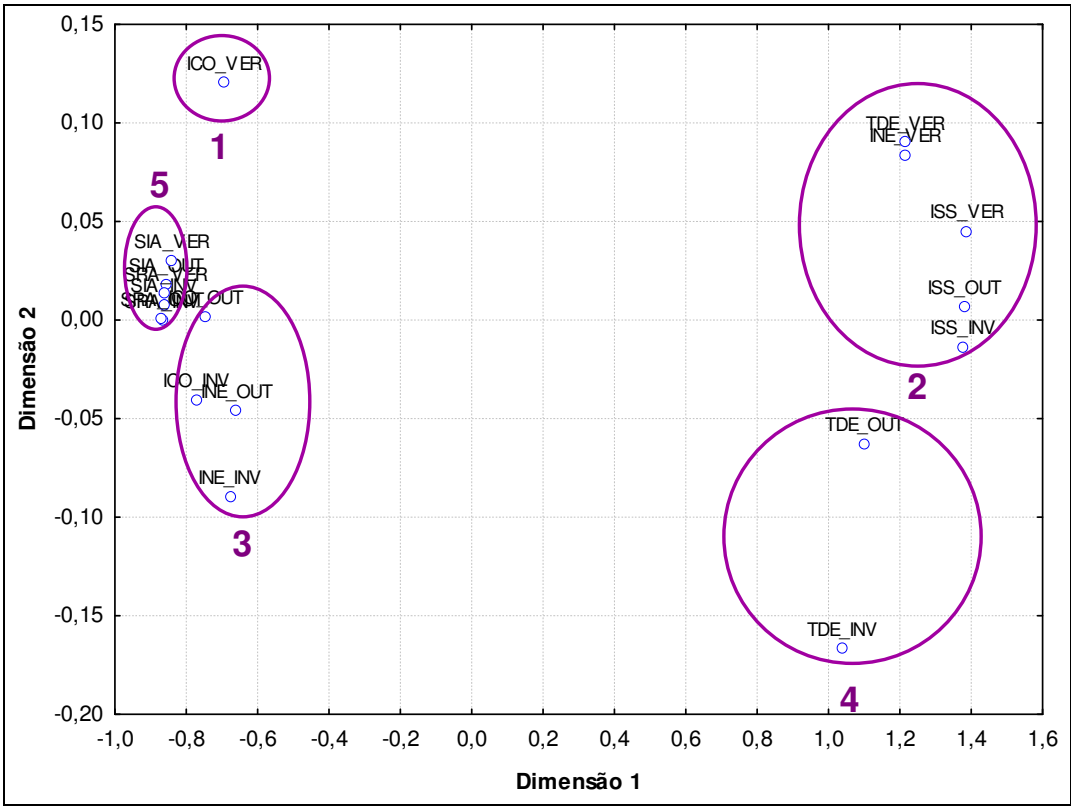

Figura 42 –Vetores de Características dos Perfis de Consumo

Analisando H_B

O método de extração de características descrito é validado mediante a exploração de um conceito derivado da análise de variância (ANOVA): *a variabilidade total do sistema*. Como discutido no Capítulo 1, existem muitas variáveis disponíveis que podem estar associadas à carga elétrica, mas em geral nem todas possuem relevância preditiva. A Equação 68 é uma identidade fundamental da análise de variância, que mostra como

um modelo de estimação explica a variabilidade de um sistema (Montgomery et al., 2004). Nessa equação, y_i é a saída real do sistema (observada), $\hat{y}_i$ é a resposta do modelo (estimada) e $\overline{y}$ é média das saídas reais.

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

Equação 68 – Variabilidade Total de um Sistema

A Equação 69, obtida a partir da Equação 68, mostra como a variabilidade total do sistema $\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)$ depende da variabilidade explicada pelo estimador $\left(\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2\right)$ e de um erro associado $\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)$. Para que um estimador seja considerado satisfatório, a variabilidade explicada por ele deve ser próxima à variabilidade total do sistema, sem o quê o erro torna-se inaceitável.

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Equação 69 – Identidade Fundamental da Análise de Variância

Um estimador neuronal ou SVM implementa uma transformação não-linear que associa as variáveis de entrada à saída. Se, num primeiro momento, todas as variáveis de entrada disponíveis são utilizadas (ou seja, o estimador é leigo), o mapeamento tende a reforçar a influência das variáveis relevantes em detrimento das demais.

Como salientado no Capítulo 1, a utilização de todas as variáveis pode degradar o desempenho do modelo, o que é normalmente percebido como um ruído na resposta. Entretanto, a amplitude desse ruído diminui durante a convergência do modelo e a contribuição das variáveis irrelevantes torna-se menor do que a das variáveis relevantes. Isso é uma decorrência do princípio de aprendizado Hebbiano, segundo o qual atividades correlacionadas reforçam sinapses ao passo que atividades não-correlacionadas as enfraquecem (Haykin, 1994). Dito de outra forma, o estimador leigo captura o grau de relevância das variáveis de entrada.

Do exposto, decorre que a relevância das variáveis de entrada pode ser percebida como o impacto que elas têm na variabilidade total da carga elétrica. Isso significa que, quanto maior a relevância de uma variável, mais ela afeta a variabilidade da carga, ana-

logamente ao que ocorre na Análise de Componentes Principais, onde a proporção da variabilidade devido a um fator principal determina a relevância deste fator (Johnson e Wichern, 2005).

Estabelecida essa relação entre a variabilidade das entradas e a da carga elétrica, torna-se possível extrair o grau de relevância de cada variável através de um teste de relevância simples, mostrado na Figura 43.

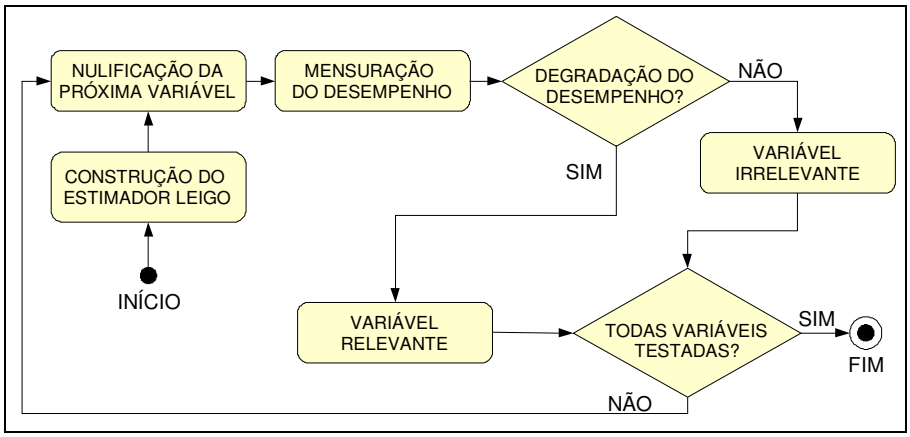

Figura 43 – Determinação da Relevância Preditiva das Variáveis Disponíveis em um Perfil

O teste de relevância proposto na Figura 43 se assemelha ao Descritor de Comprimento Mínimo (MDL, *Minimum Descritor Length*), utilizado justamente para mensurar a relevância preditiva (Taft et al, 2005). O MDL constrói um modelo de predição simples a partir de cada variável disponível. Uma vez criados, esses modelos são comparados e ordenados de acordo com a sua simplicidade algorítmica e com o grau de compressão obtido. Esta estratégia, que premia a simplicidade em detrimento da complexidade, é adotada para evitar o super ajuste (over-fitting) (Grunwald, 2005).

A Equação 70 mostra a complexidade algorítmica (complexidade de Kolmogorov) de um modelo MDL, denotada por $\mathrm{K}(h,D)$, onde h é o modelo, D é a carga elétrica e Dh é a carga elétrica como descrita por h. O objetivo é validar os modelos onde cujas complexidades sejam baixas em relação aos demais, permitindo identificar quais variáveis possuem alto potencial preditivo (Duda, 2001).

$$\mathrm{K}(h,D)=\mathrm{K}(h)+\mathrm{K}(D_h)$$

Equação 70 – Complexidade de Kolmogorov de Acordo com o MDL

Nesta pesquisa, o MDL produziu uma classificação consistente das variáveis, identificando quais eram relevantes e quais eram irrelevantes. Entretanto, para validar H_B, não é suficiente saber se a variável testada é relevante ou não, mas sim o quão relevante uma variável é. Embora o MDL atribua graus de relevância para cada variável (Taft et al, 2005), a análise de agrupamentos conduzida com estes valores não gerou os agrupamentos esperados. A conclusão que se chegou é que o MDL não fornece a precisão requerida para este tipo de análise.

Para contornar este fato e validar H_B, a concepção do MDL foi alterada, utilizando o procedimento mostrado na Figura 43. Ao invés de criar um preditor para cada variável disponível, a idéia é construir um único estimador leigo e verificar seu desempenho quando a variância de cada uma de suas entradas é anulada, uma por vez. Anular a variância significa cancelar a dispersão em torno da média, atribuindo um valor constante a cada variável. Se, nessas condições, o desempenho do modelo se degradar significativamente, a variável testada é relevante; caso contrário, é irrelevante. De toda forma, o grau de relevância dessa variável fica associado a um número real – o desempenho assim obtido.

Ora, as variáveis de entrada são comuns a todas as regiões de consumo. Uma vez que o teste descrito atribui um fator de relevância preditiva a cada variável, é possível formar um espaço geométrico, denominado *espaço causal*, onde os eixos coordenados representam a relevância de cada variável. Quando representadas neste espaço, os perfis de consumo cujos conjuntos de relevância preditiva sejam semelhantes tendem a se agrupar. Em outras palavras, regiões cujas variáveis apresentam graus de relevância semelhantes formam agrupamentos.

A Figura 44 é uma representação de baixa resolução (obtida por escalonamento multidimensional) do espaço causal, onde são traçados os vetores de cada perfil de consumo visualizado na Figura 37. Verifica-se a formação de agrupamentos nítidos e bem distribuídos, além de um fenômeno muito interessante: os agrupamentos formados, destacados com elipses, são os mesmos da Figura 42. Isso comprova que perfis similares de acordo com o critério estabelecido por H_A compartilham as mesmas variáveis preditoras. Diante desta evidência, se aceita H_B.

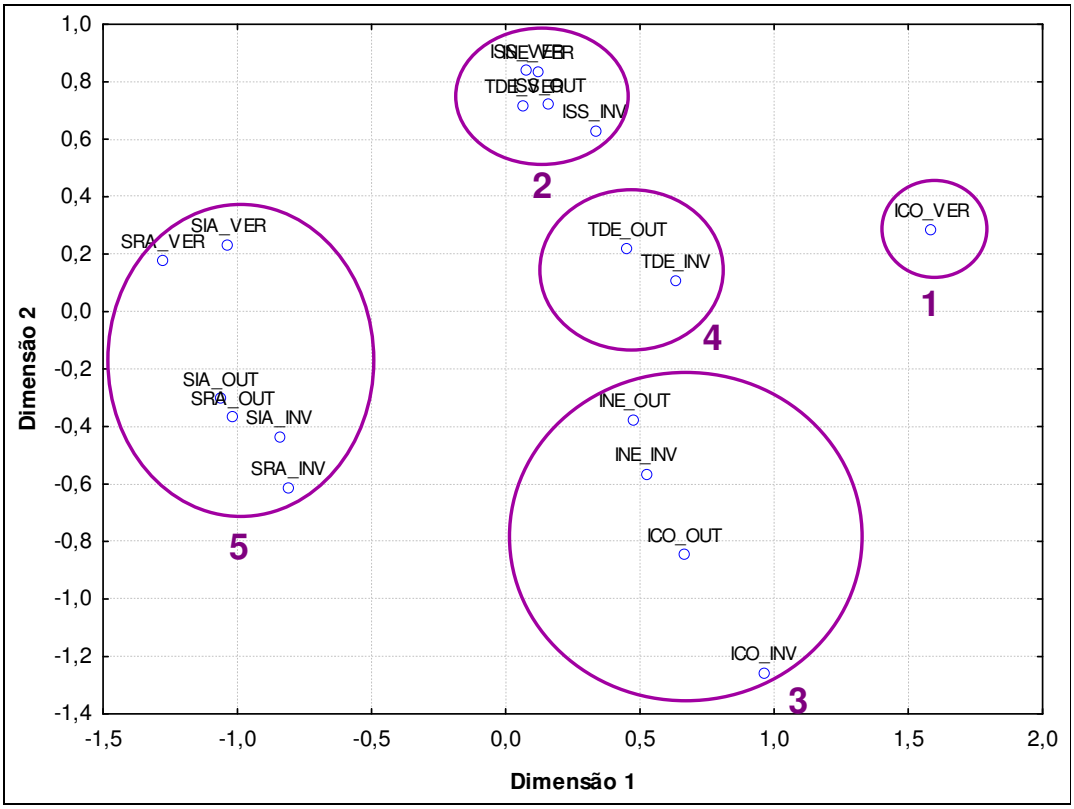

Figura 44 – Semelhança entre as Preditoras dos Perfis de Consumo

6.2 Algoritmo de classificação

A Figura 42 mostra de forma inequívoca a formação de agrupamentos, representados no plano cartesiano. Embora essa seja uma comodidade geométrica, relacionada às limitações gráficas de representação, ela mostra como a classificação pode ser feita diretamente mediante a aplicação de critérios topológicos relacionados à distância euclidiana: membros de uma classe estão mais próximos do centróide de sua classe do que de outros centróides.

Como o espaço é bastante esparso, trata-se de uma aplicação típica para técnicas de agrupamento convencionais. Neste trabalho, foi utilizado o k-means, que identificou os mesmos agrupamentos visualizados nos espaços de baixa resolução da Figura 42.

6.3 RESULTADOS

Para validar as técnicas propostas, diversos preditores neuronais puros foram criados para os perfis visualizados nas Figuras Figura 37 e Figura 42. Como o objetivo das experiências não é propor novos modelos de predição, mas sim otimizá-los, os preditores criados são básicos; isto é, não contém a sofisticação de modelos como o PCarga (Oliveira, 2004) ou o SPDS (Guo et al., 2004). Especificamente, foram criadas RNAs ajustadas através de gradiente descendente com taxa de momento e passo variável de aprendizado. A estratégia de predição para estes modelos foi a malha de atraso, como recomendado em Berry et al. (2004), Duda et al. (2001) e Haykin (1994). A Figura 45 mostra o diagrama dos preditores criados, enfatizando o emprego da malha de atraso.

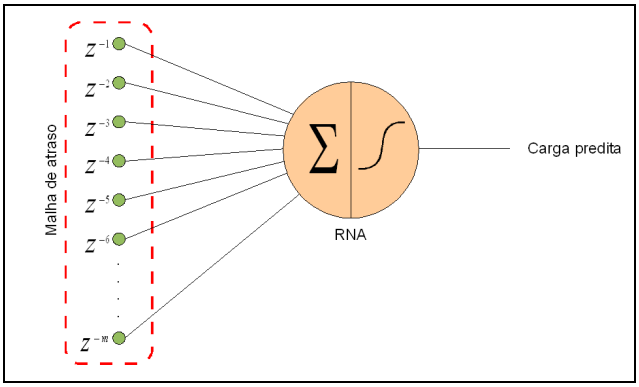

Figura 45 – Preditores Neuronais Criados para Validar a Otimização

A simplicidade dos preditores criados nestas experiências não afeta a generalidade da concepção – a meta é determinar uma regra de similaridade entre perfis de consumo diferentes. A partir daí, decorre que o conhecimento utilizado para criar o preditor de um deles pode ser utilizado para acelerar a construção de um preditor para o outro. Assim, em princípio, não é relevante considerar as estruturas preditoras utilizadas.

As simulações consideraram um histórico de 5 horas para predizer a carga meia hora adiante. Se, neste cenário, for possível demonstrar que é possível reciclar o conhecimento incorporado em preditores consolidados, espera-se obter resultados semelhantes em outras situações. Para que a otimização pudesse ser realizada, o extrator de características mostrado na Figura 41 foi implementado utilizando os mesmos estimadores leigos empregados para gerar o gráfico de desempenho da Figura 37.

Como os modelos de predição neuronais possuem um forte componente estocástico (a inicialização dos parâmetros livres), o tempo de convergência pode se alterar substancialmente de uma simulação para outra. Por esta razão, foram construídos 20 preditores para cada perfil, sendo considerados os tempos de convergência mínimo (MINTC), máximo (MAXTC) e médio (AVGTC), mostrados na Tabela 4. Inicialmente, os 20 preditores foram construídos para cada perfil sem utilizar nenhum conhecimento a priori. Dado a ausência de conhecimento prévio, a construção destes preditores demandou um custo computacional relativamente alto. Todos os vetores de características e os preditores gerados foram armazenados na Base de Conhecimento para subsidiar a construção de novos preditores.

Subseqüentemente, os perfis foram comparados uns com os outros através de seus vetores de características (Figura 42). Para um dado perfil ρ_a (coluna Perfil na Tabela 4), foi determinado o perfil ρ_b (Perfil-S) que mais lhe assemelhava através da distância Euclidiana no espaço de características. Após a comparação, a RNA que convergiu mais rápido para ρ_b foi recuperada da Base de Conhecimento e retreinada com os dados de ρ_a, produzindo uma nova RNA denominada *RNA otimizada*. Como pode ser observado na Tabela 4, a RNA otimizada convergiu mais rápido do que o tempo médio original (AVGTC) de ρ_a, como indicado pela coluna OTC (*tempo de convergência otimizado*).

Tabela 4 – Tempos de Convergência para Gerar Modelos Preditores com e sem Otimização

	Perfil	MINTC	MAXTC	AVGTC	Perfil-S	OTC	Ganho	OTC%
1	ICO_INV	615.6	1272.8	1022.9	ICO_OUT	621.5	401.4	39.2%
2	ICO_OUT	1056.5	1661.7	1350.5	ICO_INV	951.1	399.4	29.6%
3	ICO_VER	1336.4	1696.5	1530.9	ICO_OUT	1509.0	21.9	1.4%
4	INE_INV	91.6	194.5	123.9	INE_OUT	88.3	35.6	28.7%
5	INE_OUT	120.9	212.1	175.0	INE_INV	57.4	117.7	67.2%
6	INE_VER	87.7	173.8	128.2	TDE_VER	61.2	67.0	52.2%
7	ISS_INV	60.8	155.6	106.2	ISS_OUT	29.9	76.3	71.9%
8	ISS_OUT	113.4	220.0	165.1	ISS_INV	27.5	137.6	83.3%
9	ISS_VER	62.1	89.0	72.7	ISS_OUT	31.8	40.9	56.2%
10	SIA_INV	141.3	199.5	162.2	SIA_OUT	50.0	112.2	69.2%
11	SIA_OUT	129.0	185.9	156.0	SIA_INV	56.5	99.5	63.8%
12	SIA_VER	93.4	211.8	142.8	SIA_OUT	58.4	84.5	59.1%
13	SRA_INV	141.8	258.3	190.7	SRA_OUT	86.7	104.0	54.5%
14	SRA_OUT	102.4	186.5	163.6	SRA_INV	56.9	106.6	65.2%
15	SRA_VER	97.9	178.6	132.0	SIA_INV	61.5	70.4	53.4%
16	TDE_INV	84.3	139.4	106.7	TDE_OUT	59.3	47.4	44.4%
17	TDE_OUT	93.8	128.8	113.4	TDE_INV	30.7	82.7	72.9%
18	TDE_VER	91.5	341.2	153.4	INE_VER	47.7	105.7	68.9%
Médias		**251.1**	**417.0**	**333.1**		**215.9**	**117.3**	**54.51%**

Para efeitos de comparação, a Tabela 4 mostra também a relação entre o tempo médio de convergência e o TMCO (OTC%), definido na Equação 71. Como mostrado, o tempo otimizado é sempre uma fração do tempo médio de convergência, validando a solução desenvolvida e comprovando experimentalmente as hipóteses H_A e H_B.

$$OTC\% = \frac{AVGTC - OTC}{AVGTC}$$

Equação 71 – Normalização do Tempo Otimizado de Convergência

A média dos resultados obtidos foi incluída na Tabela 4. Esta agregação mostra que, em média, os tempos de convergência são sensivelmente reduzidos com o uso da otimização, comprovando a viabilidade da abordagem.

7 DISCUSSÕES E CONCLUSÃO

O Capítulo 6 mostra que a técnica de otimização desenvolvida reduz substancialmente o tempo de aprendizado dos modelos preditores gerados. O tempo exigido para criar os estimadores leigos que geram o espaço de características não foi considerado porque, uma vez criados, eles são armazenados na Base de Conhecimento e utilizados sempre que necessário. Além disso, o esforço necessário tanto para criá-los como para utilizá-los é desprezível. Desta forma, o método criado torna-se uma abordagem realística para otimizar a predição de carga de curto prazo .

O método de otimização discutido no Capítulo 6 não substitui os módulos de pré-processamento utilizados em trabalhos de predição de carga como os propostos por Guo (2004), Tao (2004), Oliveira (2004) ou Hong (2005). Este método otimiza uma estratégia de pré-processamento fornecendo dados refinados, extraídos de modelos preditores consolidados. Com isso, é possível fornecer um conjunto inicial de preditoras para o módulo de pré-processamento, ou fazer com que os novos modelos sejam construídos a partir de modelos já existentes. Em ambos os casos, o pré-processamento é iniciado a partir de um conhecimento *a priori*, restringindo o espaço de busca de uma solução ótima.

É preciso considerar que os preditores de carga utilizados na Tabela 4 (Capítulo 6) não são preditores reais, mas operam sob condições relaxadas, adotadas para viabilizar os experimentos. Isso significa que a precisão das predições é mais baixa do que a necessária para operação em campo. Embora as regularidades mostradas na Figura 42 representem tão somente os perfis de consumo, sem fazer considerações sobre a arquitetura do preditor que possa ser neles empregado, o fato é que não foram conduzidos testes com preditores reais, onde é necessário determinar a relevância preditiva das variáveis de entrada. Desta forma, a aplicação das técnicas de otimização em condições reais permanece como um aspecto em aberto, a ser explorado em trabalhos futuros.

Algumas experiências conduzidas com a Base de Conhecimento revelaram um fenômeno interessante: o grau de similaridade entre os perfis de consumo só é consistente

quando a distância entre eles no espaço de características é relativamente pequena. Caso contrário, a similaridade não será semanticamente significativa e conduzirá a conclusões espúrias. Como conseqüência, a construção de um modelo preditor só pode ser otimizada quando o perfil em questão pertencer a alguma classe (agrupamento) existente. De fato, o único agrupamento com uma única amostra (agrupamento 1, composto pelo perfil ICO_VER) apresenta um ganho muito modesto com a otimização – apenas 1.4%, como mostrado na Tabela 4. Por outro lado, os maiores ganhos foram observados nos agrupamentos onde os perfis estavam mais próximos entre si; por exemplo, o preditor de ISS_OUT, quando iniciado a partir do modelo criado para ISS_INV, apresentou um ganho de 83,3%.

O fenômeno descrito está consistente com a informação fornecida pela Figura 42, cujas categorias (destacadas por elipses) são correspondentes às categorias visualizadas na Figura 44. Entretanto, a disposição entre os agrupamentos é diferente em ambas as figuras. A título de exemplo, considere-se o caso do agrupamento 3 que na Figura 42 está mais próximo do agrupamento 5 do que na Figura 44. Nesta última figura, é o agrupamento 4 que está mais próximo do 3.

A falta de uma relação rigorosa entre o grau de similaridade e o tempo de convergência otimizado para todos os perfis evidencia uma limitação do método proposto: o espaço de características revela *tendências* e não *determinismos*, ao contrário do que uma métrica rigorosa de similaridade (a distância euclidiana) poderia talvez sugerir. Entretanto, conforme demonstrado no Capítulo 6, essas tendências são tanto mais precisas quanto menores forem as distâncias observadas entre os perfis. Ora, se a Base de Conhecimento for povoada com uma grande quantidade de dados, o espaço de características se torna menos esparso. Desta forma, a distância de um perfil desconhecido ao agrupamento mais próximo será sempre relativamente pequena e o método conduzirá a algum ganho no pré-processamento.

Os resultados empíricos obtidos indicam que a perda de precisão do critério de similaridade fora dos agrupamentos poderia ser atenuada aumentando a dimensionalidade do espaço de características. Entretanto, como a massa de dados disponível para esta pesquisa era restrita, não foi possível testar essa hipótese.

Existem outras questões relativas à extração de características. Primeiro, a extração é realizada através de *estimadores* (leigos) para otimizar a construção de *estimadores*. Ora, em princípio seria possível empregar diretamente os preditores armazenados na Base de Conhecimento para gerar as curvas de desempenho e os vetores de características. Entretanto, os preditores utilizam somente as variáveis relevantes (Tao et al., 2004; Oliveira, 2004; Hong et al., 2005). Dado que as preditoras de regiões similares podem ser ligeiramente diferentes, o uso de preditores para extrair características poderia tornar a comparação tendenciosa. De fato, a razão de ser dos estimadores leigos (descritos no Capítulo 5) é extrair as características dos perfis considerando todas as variáveis disponíveis. Como o Capítulo 6 destaca, esta estratégia produziu uma classificação coerente dos perfis de consumo.

Segundo, os gráficos obtidos nas figuras Figura 42 e Figura 44 são obtidos com estimadores leigos SVM. Gráficos análogos obtidos com estimadores neuronais não formaram padrões tão claros. Isto é intrínseco às RNAs, que possuem um forte fator estocástico associado (a inicialização dos parâmetros livres), a despeito de uma regra de aprendizado definida rigorosamente (o gradiente descendente e suas variações) (Haykin, 1998). Por outro lado, o SVM calcula o risco máximo (erro) esperado no conjunto de treinamento e limita-o tanto quanto possível (Schlkopf et al., 2001). Isto resulta em um poder de generalização maior e define padrões semanticamente consistentes como os mostrados nas figuras Figura 42 e Figura 44.

8 REFERÊNCIAS

[1] Meyer, 1971 - Meyer, H. W. *A HISTORY OF ELECTRICITY AND MAGNETISM.* ISBN: 026213070X. The MIT Press, 1971

[2] Steinbruch, 1987 et al. - Steinbruch, A., Winterle, P. *ÁLGEBRA LINEAR*. ISBN: 0074504126. Makron; 2a. edição, 1987

[3] Makridakis et al., 1997 - Makridakis, S. G., Wheelwright, S. C., Hyndman, R. J. *FORECASTING: METHODS AND APPLICATIONS.* ISBN: 0471532339. Wiley; 3 edition, 1997

[4] Haykin, 1998 - Haykin, S. *NEURAL NETWORKS.* ISBN: 0132733501 Prentice Hall; 2nd edition, 1998

[5] Vapnik, 1998 - Vapnik, V. N. *STATISTICAL LEARNING THEORY* ISBN: 0471030031. Wiley-Interscience; 1998

[6] Oppenheim et al., 1999 - Oppenheim, A. V., Schafer, R. W., Buck, J. R. *DISCRETE-TIME SIGNAL PROCESSING.* ISBN: 0137549202. Prentice Hall; 2nd edition, 1999

[7] Vapnik, 1999 - Vapnik, V. N. *THE NATURE OF STATISTICAL LEARNING THEORY.* ISBN: 0387987800. Springer; 2nd edition, 1999

[8] Duda et al., 2000 - Duda, R. O., Hart, P. E., Stork, D. G. *PATTERN CLASSIFICATION.* ISBN: 0471056693. Wiley-Interscience; 2nd edition, 2000

[9] Cristiani, 2001 - Cristianini, N., Shawe-Taylor, J. *AN INTRODUCTION TO SUPPORT VECTOR MACHINES AND OTHER KERNEL-BASED LEARNING METHODS.* ISBN: 0521780195. Cambridge University Press; 2000

[10] Schlkopf et al., 2001 - Schlkopf, B., Smola, A. *LEARNING WITH KERNELS - SUPPORT VECTOR MACHINES, REGULARIZATION, OPTIMIZATION, AND BEYOND*. ISBN: 0262194759. MIT Press; 2001

[11] Montgomery et al., 2001 - Montgomery, D. C., Peck, E. A., Vining, G. *INTRODUCTION TO LINEAR REGRESSION ANALYSIS.* ISBN: 0471315656. Wiley-Interscience; 3rd edition, 2001

[12] Herbrich, 2001 - Herbrich, R. *LEARNING KERNEL CLASSIFIERS: THEORY AND ALGORITHMS.* ISBN: 026208306X. MIT Press, 2001

[13] Hand et al., 2001 - Hand, D. J., Mannila, H., Smyth, P. *PRINCIPLES OF DATA MINING*. ISBN: 026208290X. MIT Press, 2001

[14] Johnson et al., 2002 - Johnson, R. A., Wichern, D. W. *APPLIED MULTIVARIATE STATISTICAL ANALYSIS*. ISBN: 0130925535. Prentice Hall;

5th edition, 2002

[15] Acha et al., 2002 - Acha, E., Agelidis, V., Anaya, O., Miller, T. J. E. *POWER ELECTRONIC CONTROL IN ELECTRICAL SYSTEMS*. ISBN: 0750651261. Newnes; 2002

[16] Rencher, 2002 - Rencher, A. C. *METHODS OF MULTIVARIATE ANALYSIS*. ISBN: 0471418897. John Wiley & Sons, 2002

[17] Santoso et al., 2002 Santoso, S., Beaty, H. W., Dugan, R. C., McGranaghan, M. M. *ELECTRICAL POWER SYSTEMS QUALITY*. ISBN: 007138622X. McGraw-Hill Professional; 2 edition, 2002

[18] Hastie et al., 2003 - Hastie, T., Tibshirani, R., Friedman, J. H. *THE ELEMENTS OF STATISTICAL LEARNING*. ISBN: 0387952845. Springer; 2003

[19] Iyer et al., 2003 - Iyer, V., Fung, C.C., Gedeon, T. *A FUZZY NEURAL APPROACH TO ELECTRICITY LOAD AND SPOT-PRICE FORECASTING IN A DEREGULATED ELECTRICITY MARKET*.TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, Volume: 4, On page(s): 1479-1482 Vol.4 ISBN: 0-7803-8162-9; Oct. 2003

[20] Halliday et al., 2004 - Halliday, D., Resnick, R., Walker, J. *FUNDAMENTALS OF PHYSICS*. ISBN: 0471216437. Wiley; 7th edition, 2004

[21] Tao et al., 2004 - Tao, X., Renmu, H., Peng, W., Dongjie, X. *INPUT DIMENSION REDUCTION FOR LOAD FORECASTING BASED ON SUPPORT VECTOR MACHINES*. Electric Utility Deregulation, Restructuring and Power Technologies, 2004. (DRPT 2004). Proceedings of the 2004 IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies, Volume 2, On Page(s): 510 - 514 Vol.2; April 2004

[22] Guo et al., 2004 - Guo, X., Chen, Z., Ge, H., Liang, Y. *SHORT-TERM LOAD FORECASTING USING NEURAL NETWORK WITH PRINCIPAL COMPONENTS ANALYSIS*. Proceedings of the Third International Conference on Machine Learning and Cybernetics, Volume 6, On Page(s): 3365 - 3369 vol.6; Aug. 2004

[23] Oliveira, 2004 - Oliveira, C. M. *MODELO ADAPTATIVO PARA PREVISÃO DE CARGA ATIVA DE CURTO PRAZO*. Tese de Doutorado, Depto. de Engenharia de Produção – Universidade Federal de Santa Catarina, 2004

[24] Pindyck et al., 2004 - Pindyck, R. S., Rubinfield, D. L. *ECONOMETRIA: MODELOS & PREVISÕES*. ISBN: 8535213430. Campus; 4ª Edição, 2004

[25] Berry et al., 2004 - Berry, M. J. A, Linoff G. S. *DATA MINING TECHNIQUES: FOR MARKETING, SALES, AND CUSTOMER RELATIONSHIP MANAGEMENT*. ISBN: 0471470643. Wiley Publishing; 2nd edition, 2004

[26] Pansini, 2005 - Pansini, A. J. *GUIDE TO ELECTRICAL POWER DISTRIBUTION SYSTEMS*. ISBN: 084933666X. CRC; 6th edition, 2005

[27] Grunwald, 2005 - Grunwald, P. D. (ed.), Myung, I. J.(ed.), Pitt, M. A. (ed.): *ADVANCES IN MINIMUM DESCRIPTION LENGTH: THEORY AND APPLICATIONS*. ISBN: 0262072629. MIT; 2005

[28] Taft et al., 2005 - Taft, M., Krishnan R., Hornick, M., Muhkin, D., Tang, G., Thomas, S., Stengard, P. *ORACLE DATA MINING CONCEPTS*. Part no: B14339-01. Oracle Press; 2005

[29] Hong et al., 2005 - Hong, W., Pai, P., Chen, C., Lin, C. *ELECTRICITY LOAD FORECASTING BY USING SUPPORT VECTOR MACHINES WITH SIMULATED ANNEALING ALGORITHM*. Proceedings on the 17th IMACS World Congress Scientific Computation, Applied Mathematics and Simulation, Paris, France; July 2005

[30] Niu et al., 2005 - Niu, D., Wang, Q., Li, J. *SHORT TERM LOAD FORECASTING MODEL USING SUPPORT VECTOR MACHINE BASED ON ARTIFICIAL NEURAL NETWORK*. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Volume 7, On Page(s): 4260 - 4265 Vol. 7; Aug. 2005

[31] Guo et al., 2006 Guo, Y., Niu, D., Chen,Y. *SUPPORT VECTOR MACHINE MODEL IN ELECTRICITY LOAD FORECASTING.* Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August; Aug. 2006

9 ANEXOS

9.1 Variáveis disponíveis para a predição de carga

Os tempos dados em minutos se referem ao horário em que a medição for efetuada, por exemplo:

- `Derivada 2ª. da Vel. Vento (30 minutos)`: derivada segunda da velocidade do vento há 30 minutos atrás;
- `Derivada Temp. (60 minutos, Ano Anterior)`: derivada da temperatura há 60 minutos atrás no ano anterior;
- `Derivada 2ª. Temp. (Atual, Semana Anterior)`: derivada de segunda ordem da temperatura no mesmo horário da semana anterior.

Índice	Variável	Descrição
1	AGORA_DER2_PRS_0MINUTOS	Derivada 2ª. da Vel. Vento (Atual)
2	AGORA_DER2_PRS_30MINUTOS	Derivada 2ª. da Vel. Vento (30 minutos)
3	AGORA_DER2_TCW_30MINUTOS	Derivada 2ª. da Carga (30 minutos)
4	AGORA_DER2_TMP_0MINUTOS	Derivada 2ª. da Temperatura (Atual)
5	AGORA_DER2_TMP_30MINUTOS	Derivada 2ª. da Temperatura (30minutos)
6	AGORA_DER_PRS_0MINUTOS	Derivada da Vel. Vento (Atual)
7	AGORA_DER_PRS_30MINUTOS	Derivada da Vel. Vento (30 minutos)
8	AGORA_DER_PRS_60MINUTOS	Derivada da Vel. Vento (60 minutos)
9	AGORA_DER_TCW_30MINUTOS	Derivada da Carga (30 minutos)
10	AGORA_DER_TCW_60MINUTOS	Derivada da Carga (60 minutos)
11	AGORA_DER_TMP_0MINUTOS	Derivada da Temperatura (Atual)
12	AGORA_DER_TMP_30MINUTOS	Derivada da Temperatura (30 minutos)
13	AGORA_DER_TMP_60MINUTOS	Derivada da Temperatura (60 minutos)
14	AGORA_PRS_0MINUTOS	Velocidade do Vento (Atual)
15	AGORA_PRS_30MINUTOS	Velocidade do Vento (30 minutos)
16	AGORA_PRS_60MINUTOS	Velocidade do Vento (60 minutos)
17	AGORA_PRS_90MINUTOS	Velocidade do Vento (90 minutos)
18	AGORA_PRS_MAX	Velocidade do Vento (Máxima)
19	AGORA_PRS_MED	Velocidade do Vento (Média)
20	AGORA_PRS_MIN	Velocidade do Vento (Mínima)
21	AGORA_TCW_30MINUTOS	Carga (30 minutos)
22	AGORA_TCW_60MINUTOS	Carga (60 minutos)
23	AGORA_TCW_90MINUTOS	Carga (90 minutos)
24	AGORA_TMP_0MINUTOS	Temperatura (Atual)
25	AGORA_TMP_30MINUTOS	Temperatura (30 minutos)
26	AGORA_TMP_60MINUTOS	Temperatura (60 minutos)
27	AGORA_TMP_90MINUTOS	Temperatura (90 minutos)
28	AGORA_TMP_MAX	Temperatura (Máxima)
29	AGORA_TMP_MED	Temperatura (Média)
30	AGORA_TMP_MIN	Temperatura (Mínima)
31	ANO_DER2_PRS_0MINUTOS	Derivada 2ª. Vel. Vento (Atual, Ano Anterior)
32	ANO_DER2_PRS_30MINUTOS	Derivada 2ª. Vel. Vento (30 minutos, Ano Anterior)

33	ANO_DER2_TCW_0MINUTOS	Derivada 2ª. Carga (Atual, Ano Anterior)
34	ANO_DER2_TCW_30MINUTOS	Derivada 2ª. Carga (30 minutos, Ano Anterior)
35	ANO_DER2_TMP_0MINUTOS	Derivada 2ª. Temp. (Atual, Ano Anterior)
36	ANO_DER2_TMP_30MINUTOS	Derivada 2ª. Temp. (30 minutos, Ano Anterior)
37	ANO_DER_PRS_0MINUTOS	Derivada 2ª. Vel. Vento (Atual, Ano Anterior)
38	ANO_DER_PRS_30MINUTOS	Derivada 2ª. Vel. Vento (30 minutos, Ano Anterior)
39	ANO_DER_PRS_60MINUTOS	Derivada 2ª. Vel. Vento (60 minutos, Ano Anterior)
40	ANO_DER_TCW_0MINUTOS	Derivada Carga (Atual, Ano Anterior)
41	ANO_DER_TCW_30MINUTOS	Derivada Carga (30 minutos, Ano Anterior)
42	ANO_DER_TCW_60MINUTOS	Derivada Carga (60 minutos, Ano Anterior)
43	ANO_DER_TMP_0MINUTOS	Derivada Temp. (Atual, Ano Anterior)
44	ANO_DER_TMP_30MINUTOS	Derivada Temp. (30 minutos, Ano Anterior)
45	ANO_DER_TMP_60MINUTOS	Derivada Temp. (60 minutos, Ano Anterior)
46	ANO_PRS_0MINUTOS	Velocidade do Vento (Atual, Ano Anterior)
47	ANO_PRS_30MINUTOS	Velocidade do Vento (30 minutos, Ano Anterior)
48	ANO_PRS_60MINUTOS	Velocidade do Vento (60 minutos, Ano Anterior)
49	ANO_PRS_90MINUTOS	Velocidade do Vento (90 minutos, Ano Anterior)
50	ANO_PRS_MAX	Velocidade do Vento (Máxima, Ano Anterior)
51	ANO_PRS_MED	Velocidade do Vento (Média, Ano Anterior)
52	ANO_PRS_MIN	Velocidade do Vento (Mínima, Ano Anterior)
53	ANO_TCW_0MINUTOS	Carga (Atual, Ano Anterior)
54	ANO_TCW_30MINUTOS	Carga (30 minutos, Ano Anterior)
55	ANO_TCW_60MINUTOS	Carga (60 minutos, Ano Anterior)
56	ANO_TCW_90MINUTOS	Carga (90 minutos, Ano Anterior)
57	ANO_TCW_MAX	Carga (Máxima, Ano Anterior)
58	ANO_TCW_MED	Carga (Média, Ano Anterior)
59	ANO_TCW_MIN	Carga (Mínima, Ano Anterior)
60	ANO_TMP_0MINUTOS	Temperatura (Atual, Ano Anterior)
61	ANO_TMP_30MINUTOS	Temperatura (30 minutos, Ano Anterior)
62	ANO_TMP_60MINUTOS	Temperatura (60 minutos, Ano Anterior)
63	ANO_TMP_90MINUTOS	Temperatura (90 minutos, Ano Anterior)
64	ANO_TMP_MAX	Temperatura (Máxima, Ano Anterior)
65	ANO_TMP_MED	Temperatura (Média, Ano Anterior)
66	ANO_TMP_MIN	Temperatura (Mínima, Ano Anterior)
67	DIA_MES	Dia do mês
68	DIA_SEMANA	Dia da Semana
69	DIA_SEMANA_COS	
70	DIA_SEMANA_SEN	
71	HORA_MINUTO	Hora do dia
72	MEIA_HORA_DIA_COS	
73	MEIA_HORA_DIA_SEN	
74	MES	Mês
75	SEMANA_ANO_COS	
76	SEMANA_ANO_SEN	
77	SEMANA_DER2_PRS_0MINUTOS	Derivada 2ª. Vel. Vento (Atual, Semana Anterior)
78	SEMANA_DER2_PRS_30MINUTOS	Derivada 2ª. Vel. Vento (30 minutos, Semana Anterior)
79	SEMANA_DER2_TCW_0MINUTOS	Derivada 2ª. Carga (Atual, Semana Anterior)
80	SEMANA_DER2_TCW_30MINUTOS	Derivada 2ª. Carga (30 minutos, Semana Anterior)
81	SEMANA_DER2_TMP_0MINUTOS	Derivada 2ª. Temp. (Atual, Semana Anterior)

82	SEMANA_DER2_TMP_30MINUTOS	Derivada 2ª. Temp. (30 minutos, Semana Anterior)
83	SEMANA_DER_PRS_0MINUTOS	Derivada Vel. Vento (Atual, Semana Anterior)
84	SEMANA_DER_PRS_30MINUTOS	Derivada Vel. Vento (30 minutos, Semana Anterior)
85	SEMANA_DER_PRS_60MINUTOS	Derivada Vel. Vento (60 minutos, Semana Anterior)
86	SEMANA_DER_TCW_0MINUTOS	Derivada Carga (Atual, Semana Anterior)
87	SEMANA_DER_TCW_30MINUTOS	Derivada Carga (30 minutos, Semana Anterior)
88	SEMANA_DER_TCW_60MINUTOS	Derivada Carga (60 minutos, Semana Anterior)
89	SEMANA_DER_TMP_0MINUTOS	Derivada Temp. (Atual, Semana Anterior)
90	SEMANA_DER_TMP_30MINUTOS	Derivada Temp. (30 minutos, Semana Anterior)
91	SEMANA_DER_TMP_60MINUTOS	Derivada Temp. (60 minutos, Semana Anterior)
92	SEMANA_PRS_0MINUTOS	Velocidade do Vento (Atual, Semana Anterior)
93	SEMANA_PRS_30MINUTOS	Velocidade do Vento (30 minutos, Semana Anterior)
94	SEMANA_PRS_60MINUTOS	Velocidade do Vento (60 minutos, Semana Anterior)
95	SEMANA_PRS_90MINUTOS	Velocidade do Vento (90 minutos, Semana Anterior)
96	SEMANA_PRS_MAX	Velocidade do Vento (Máxima, Semana Anterior)
97	SEMANA_PRS_MED	Velocidade do Vento (Média, Semana Anterior)
98	SEMANA_PRS_MIN	Velocidade do Vento (Mínima, Semana Anterior)
99	SEMANA_TCW_0MINUTOS	Carga (Atual, Semana Anterior)
100	SEMANA_TCW_30MINUTOS	Carga (30 minutos, Semana Anterior)
101	SEMANA_TCW_60MINUTOS	Carga (60 minutos, Semana Anterior)
102	SEMANA_TCW_90MINUTOS	Carga (90 minutos, Semana Anterior)
103	SEMANA_TCW_MAX	Carga (Máxima, Semana Anterior)
104	SEMANA_TCW_MED	Carga (Média, Semana Anterior)
105	SEMANA_TCW_MIN	Carga (Mínima, Semana Anterior)
106	SEMANA_TMP_0MINUTOS	Temperatura (Atual, Semana Anterior)
107	SEMANA_TMP_30MINUTOS	Temperatura (30 minutos, Semana Anterior)
108	SEMANA_TMP_60MINUTOS	Temperatura (60 minutos, Semana Anterior)
109	SEMANA_TMP_90MINUTOS	Temperatura (90 minutos, Semana Anterior)
110	SEMANA_TMP_MAX	Temperatura (Máxima, Semana Anterior)
111	SEMANA_TMP_MED	Temperatura (Média, Semana Anterior)
112	SEMANA_TMP_MIN	Temperatura (Mínima, Semana Anterior)

9.2 Variáveis relevantes em cada perfil de consumo

O mesmo método utilizado para gerar o espaço causal (Capítulo 6) pode ser usado para testar a relevância das variáveis em cada perfil de consumo. A tabela a seguir mostra quais são as variáveis consideradas mais relevantes em cada perfil, de acordo com o método descrito no Capítulo 6.

Perfil de consumo	**Variáveis relevantes**
ICO_INVER	009, 010, 018, 019, 020, 021, 040, 041, 046, 047, 048, 049, 050, 051, 052, 074, 086, 087, 088, 096, 097, 098, 099, 100
ICO_OUTON	009, 010, 014, 015, 016, 017, 018, 019, 020, 021, 022, 023, 040, 041, 046, 047, 048, 049, 050, 051, 052, 076, 086, 087, 088, 092, 093, 094, 095, 096, 097, 098, 099
ICO_VERAO	009, 010, 014, 015, 016, 017, 018, 019, 020, 021, 022, 023, 024, 025, 026, 027, 028, 029, 030, 040, 041, 046, 047, 048, 049, 050, 051, 052, 053, 062, 063, 064, 065, 066, 068, 075, 076, 086, 092, 093, 094, 095, 096, 097, 098, 106, 107, 108, 109, 110, 111, 112
INE_INVER	009, 010, 017, 018, 019, 020, 021, 022, 023, 040, 050, 051, 052, 053, 074, 086, 087, 096, 097, 098, 099, 100, 101, 102
INE_OUTON	009, 010, 018, 019, 021, 022, 023, 040, 041, 050, 051, 052, 053, 054, 055, 056, 076, 086, 087, 098, 099
INE_VERAO	009, 021, 022, 023
ISS_INVER	021, 022, 023
ISS_OUTON	021, 022, 023
ISS_VERAO	021, 022, 023
SIA_INVER	009, 010, 021, 022, 023, 086
SIA_OUTON	009, 010, 021, 022, 023, 040, 053, 068, 076, 086
SIA_VERAO	009, 010, 021, 022, 023, 068, 075, 076, 086, 087, 110, 112
SRA_INVER	003, 009, 010, 021, 022, 040, 086, 087
SRA_OUTON	009, 010, 021, 022, 023, 040, 068, 086, 087
SRA_VERAO	009, 010, 021, 022, 023, 028, 068, 075, 086, 099, 110, 111
TDE_INVER	009, 010, 021, 022, 023, 028, 068, 075, 086, 099, 110, 111
TDE_OUTON	009, 010, 021, 022, 023, 053, 054, 086, 099, 100
TDE_VERAO	009, 010, 021, 022, 023, 040, 053, 054, 055, 086, 099